February 2022

Electricity Prices Predictions in Spain. Data Science Master's Dissertation.

KSCHOOL

PAULA CERVILLA GARCÍA

Table of Contents

Introduction

The aim of this project was to predict the overall daily electricity price in Spain, after deeply studying and understanding all the relevant variables that make up the mentioned price.

Historical electricity prices datasets, as well as weather conditions and Brent crude oil futures prices were used along with forecasting/prediction models and various techniques of data mining, engineering and processing.

This topic was chosen because it became an issue of real interest in 2021, due to the enormous and non-expected price increases through the year.

As it can be seen from the code and analysis of the data, 2021 prices have reached maximum levels never seen before in the market and, therefore, very difficult to predict.

Another reason why this topic was chosen, was the availability of data. Total electricity prices, as well as all of its components, are published daily. Weather variables, such as daily temperatures, precipitation and wind gusts are also available online. Furthermore, the Brent crude oil futures historical prices are also available in many web portals.

In regards to the state of art, there are not any official sources where you can check electricity prices predictions. However, there are various dissertations online from other students where this topic was analysed using different machine learning techniques, like Neural Networks.

In addition, there is a company called "AleaSoft Energy Forecasting", which offers short-, medium- and long-term forecasts for energy prices, but those forecasts cannot be checked online.

Raw Data Description

Data was obtained from different web pages, depending on the type of data.

Data could be categorised in four sections:

- Electricity price data
- Brent Oil Crude futures data
- Weather variables data
- Population data

1. Electricity Price Data

Historical electricity prices, which was the dependent variable and the price that this project wanted to predict, along with the different components of the final electricity prices in Spain were obtained from the CNMC (Comisión Nacional de los Mercados y la Competencia).

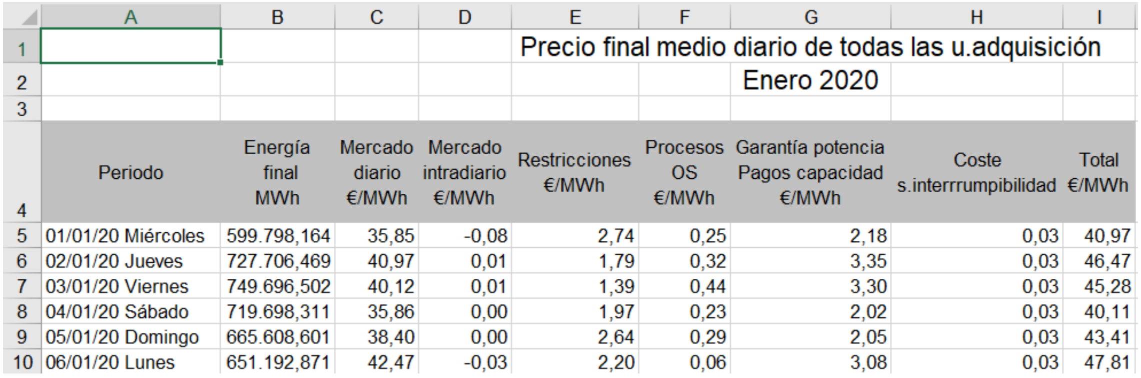
This institution "oversees the functioning and level of competition in the electricity market, within the scope of both the wholesale and retail markets, and they supervise operation of the system" and "oversees the level of effectiveness of the deregulated market and competition, in the wholesale as well as retail markets. This includes, among other actions, complaints submitted by consumers regarding electricity and natural gas, the regulated auctions for term contracting of electricity" (Comisión Nacional de los Mercados y la Competencia, 2021).

From here, we can download the market prices from the "Prices from the Electricity Production Market" section, by year. The earliest we can obtain data is from 2013. The latest data is for 2021 up to September, being all 2021 numbers provisional and not definitive, this meaning that the CNMC has not finished with its checks on the prices for all months of 2021. Some of the latest months of 2020 are also provisional prices but the difference between provisional and definitive prices has been checked and it is minimal, so it is fair to use these numbers in the models.

The structure of the downloaded data is the following:

- A zip file by year
- Within that zip file, zip files by month
- Within the zip files by month, zip files where the data is distributed by:
    - Day and hour
    - Day
    - Total of the month
- All of the above distributions are also provided for the different granularity levels of components of the final electricity price.

We are using the daily prices with all of its components, so we need to use the data within the folders called **"PFMDIASM_TOD"**. We have one Excel file by month of the year, with one row by day of the month. Below there is an example:



Precio final medio diario de todas las u.adquisición

Enero 2020

Periodo	Energía final MWh	Mercado diario €/MWh	Mercado intradiario €/MWh	Restricciones €/MWh	Procesos OS €/MWh	Garantía potencia Pagos capacidad €/MWh	Coste s.interrrumpibilidad	Total €/MWh
01/01/20 Miércoles	599.798,164	35,85	-0,08	2,74	0,25	2,18	0,03	40,97
02/01/20 Jueves	727.706,469	40,97	0,01	1,79	0,32	3,35	0,03	46,47
03/01/20 Viernes	749.696,502	40,12	0,01	1,39	0,44	3,30	0,03	45,28
04/01/20 Sábado	719.698,311	35,86	0,00	1,97	0,23	2,02	0,03	40,11
05/01/20 Domingo	665.608,601	38,40	0,00	2,64	0,29	2,05	0,03	43,41
06/01/20 Lunes	651.192,871	42,47	-0,03	2,20	0,06	3,08	0,03	47,81

Figure 1 – Example of Raw Data Electricity Prices

A data dictionary with all fields can be seen below:

Field	Meaning
Periodo	Date
Energía final MWh	Final energy that was negotiated for each day measured in MWh
Mercado diario €/MWh	Price of each MWh in the day-ahead session in EUR
Mercado intradiario €/MWh	Price of each MWh in the Intraday session in EUR
Restricciones €/MWh	Funding of the technical restrictions in the resolution process measured in EUR by MWh
Procesos OS €/MWh	The amount corresponding to the deviations produced and the adjustment services provided measured in EUR by MWh
Garantía potencia Pagos capacidad €/MWh	Amount charged for the financing of the power warranty service measured in EUR by MWh
Coste s.interrumpibilidad	Amount charged for the financing of the interruptibility service measured in EUR by MWh
Total €/MWh	Average final price of energy measured in EUR by MWh

Table 1 – Data Dictionary Electricity Prices

In order to understand how the electricity market works in Spain a bit more, below is a simple explanation from OMIE, "the nominated electricity market operator (NEMO) for managing the Iberian Peninsula's day-ahead and intraday electricity markets" (OMIE, 2021):

"The electricity market is structured into a Day-ahead market, an Intraday auction market and an Intraday continuous market" (OMIE, 2021). For more detailed information, please visit the following link: https://www.omie.es/en/mercado-de-electricidad

2. Brent Oil Crude Futures Data

Brent Oil Crude futures prices are an important determinant of the final electricity prices, as well as for example, the share of production generated by renewable energy and the price of carbon emissions.

Therefore, historical prices from 2013 to 2021 were downloaded from Investing.com: https://es.investing.com/commodities/brent-oil-historical-data (Investing.com, 2021), so they could be used when modelling.

A CSV separated by commas, with headers is downloaded, containing one row per day.

A data dictionary with all fields can be seen below:

Field	Meaning
Fecha	Date
Último	Closing price of the session for the day in EUR
Apertura	Opening price of the session for the day in EUR
Máximo	Highest price of the session for the day in EUR
Mínimo	Lowest price of the session for the day in EUR
Vol.	Number of contracts that were sold each day (thousands)
% var.	Change in % between the close prices

Table 2 - Data Dictionary Brent Oil Crude Futures Prices

The Brent Oil Crude futures market, similarly to the stock markets, do not operate on the weekends and Bank Holidays, thus data was missing for those days.

Friday prices were used for Saturday and Sunday prices, as those were the latest and current prices for those days. For Bank Holidays, the previous day prices were used for the same reason.

3. Weather Variables Data

Temperatures, precipitations and wind gusts are also determinants of the final electricity prices, so those were also used as independent variables when modelling.

This data was downloaded from a webpage that compiles AEMET (Agencia Estatal de Meteorología), or the Meteorological State Agency of Spain, data historically. Data had a cost of 1.60€ and availability from 2013.

One CSV separated by ";" for each of the 850 weather stations across Spain, whose data is available from, was downloaded along with a list of all weather stations names, single identifier ID, location and some other identifiers.

Each CSV name is the single identifier ID.

A data dictionary for the weather stations list can be seen below (raw data had no headers, but these were added):

Field	**Meaning**
Identificador_est	Single weather station identifier ID
Nombre_est	Weather Station name
Municipio	Town/city where the weather station is located
Provincia	Province where the weather station is located
Altura	Altitude above the sea where the weather station is located
Longitud	Longitude of the weather station
Latitud	Latitude of the weather station

Table 3 - Data Dictionary Weather Stations List

A data dictionary for the weather stations data CSVs can be seen below:

Field	**Meaning**
Id	Single weather station identifier ID
Fecha	Date
Tmax	Maximum Temperature (ºC)
HTmax	Time Maximum Temperature
Tmin	Minimum Temperature (ºC)
HTmin	Time Minimum Temperature
Tmed	Average Temperature (ºC)
Racha	Maximum Wind Gust (Km/h)
HRacha	Time of Maximum Wind Gust
Vmax	Average Wind Speed (Km/h)
HVmax	Time of Maximum Wind Speed
TPrec	Total daily precipitation (mm)
Prec1	Precipitation from 0 to 6 hours (mm)
Prec2	Precipitation from 6 to 12 hours (mm)
Prec3	Precipitation from 12 to 18 hours (mm)
Prec4	Precipitation from 18 to 24 hours (mm)

Table 4 - Data Dictionary Weather Stations Data

Some restraints of this data were:

- Data was collected automatically and no more checks apart from the real time automatic ones were performed on it, so that means it could have minor mistakes.
- Some weather stations might not have data for all variables and some days might be missing for some weather stations.

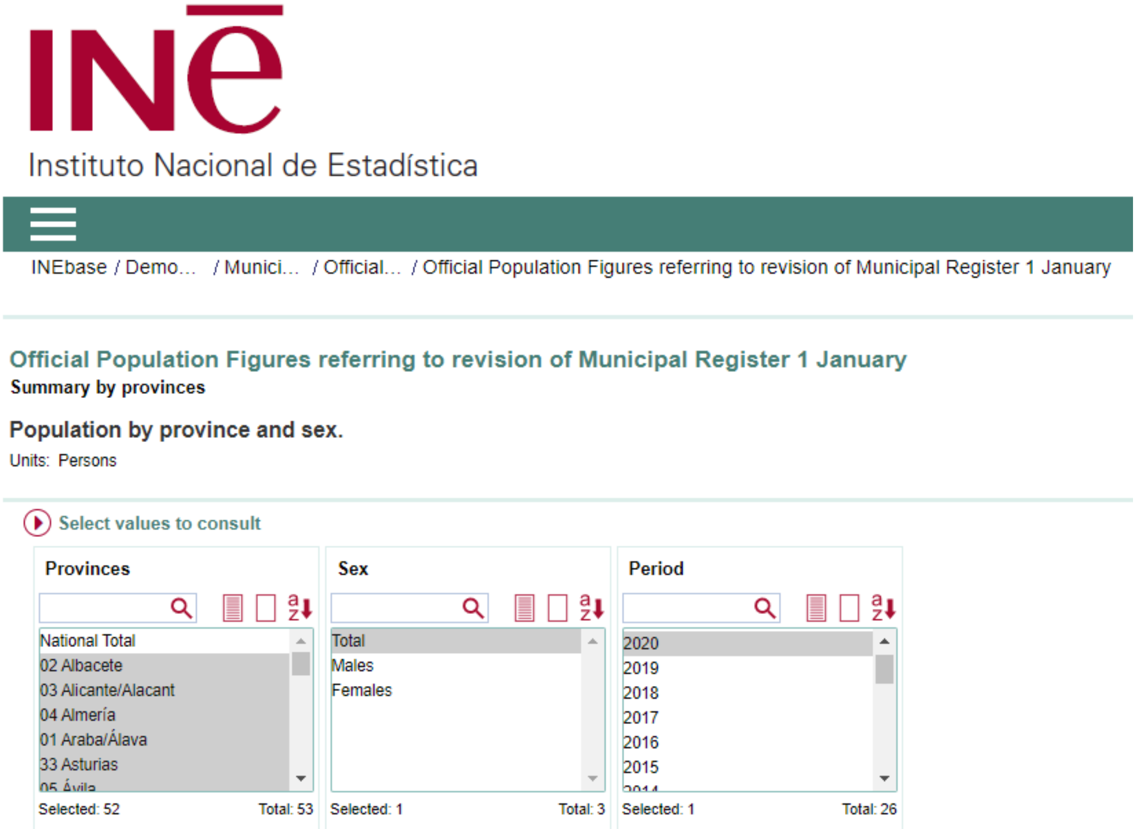
The first step performed with this data was to get all the weather stations identification codes by province. After that, a median average by province was done along with a weighted average by each province population, so the values for those provinces with more inhabitants were more important. This was done because one single value by day was needed, in order to have this data in the same format as the electricity prices and the Brent oil crude futures.

4. Population Data

In order to do the weighted average by population of each province, this data was downloaded from the Spanish National Statistics Institute (INE), in a CSV file.

Records from 2020 were used, due to availability at the time of this project.

Below, and screenshot of how the data was selected can be seen:



INE
Instituto Nacional de Estadística

INEbase / Demo… / Munici… / Official… / Official Population Figures referring to revision of Municipal Register 1 January

Official Population Figures referring to revision of Municipal Register 1 January
Summary by provinces

Population by province and sex.
Units: Persons

Select values to consult

Provinces
National Total
02 Albacete
03 Alicante/Alacant
04 Almería
01 Araba/Álava
33 Asturias
05 Ávila
Selected: 52 Total: 53

Sex
Total
Males
Females
Selected: 1 Total: 3

Period
2020
2019
2018
2017
2016
2015
2014
Selected: 1 Total: 26

Figure 2 - Screenshot INE, Population Data

A data dictionary for population raw data can be seen below:

Field	Meaning
Provincias	Number code and name of each Spanish province
Sexo	Sex of the population (Female + Male)
Periodo	Year (2020)
Total	Inhabitants

Table 5 - Data Dictionary Population Data

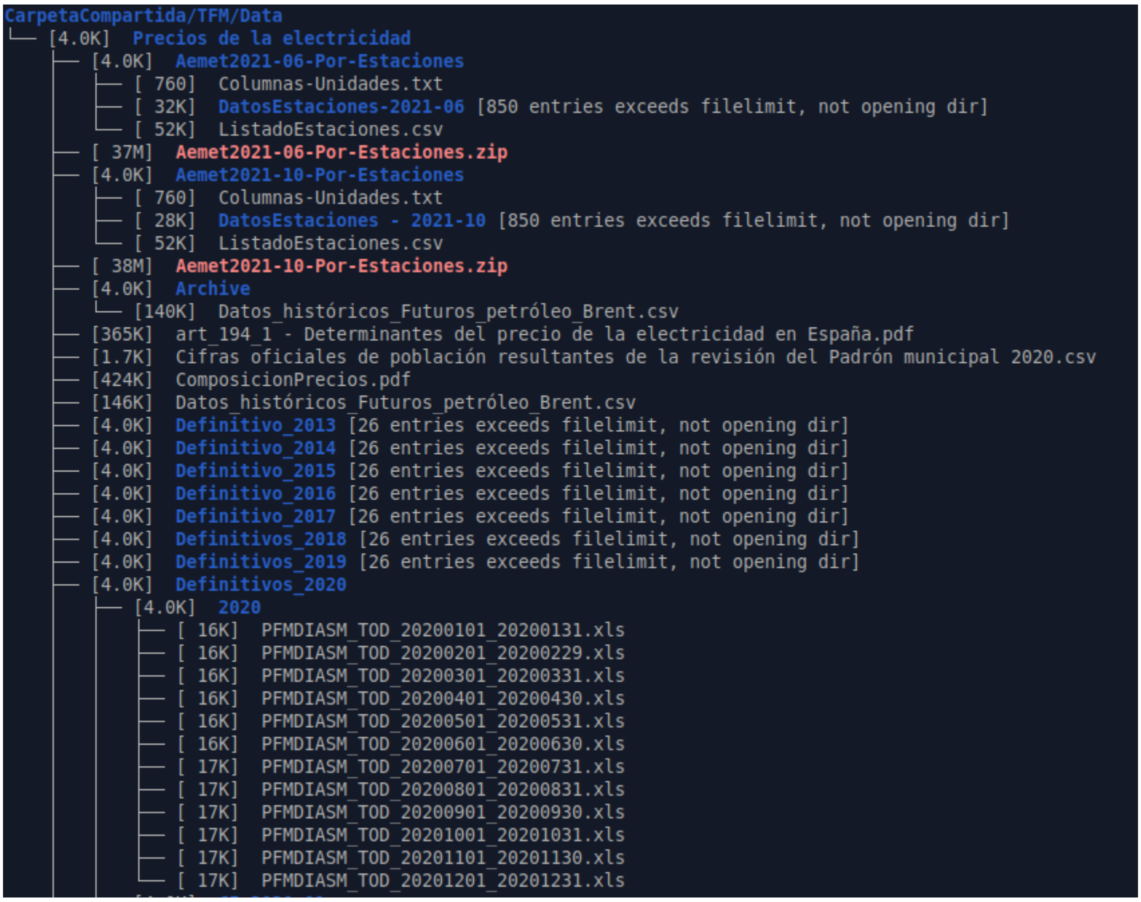
5. Folder's Organisation

To be able to run the code without issues, please place the raw data in the same structure as explained down below.

My main directory was "/home/dsc/CarpetaCompartida/TFM/", and within it, there was a folder called Data > Precios de la electricidad: "/home/dsc/CarpetaCompartida/TFM/Data/Precios de la electricidad/"

In the directory above, there were the folders with the electricity prices by year and another 2 folders with weather data (one of them until June 2021 and the other folder contains data until October 2021), along with the csv for the Brent oil crude futures prices and the population data csv.

Please see below an overview of the directories' organisation:



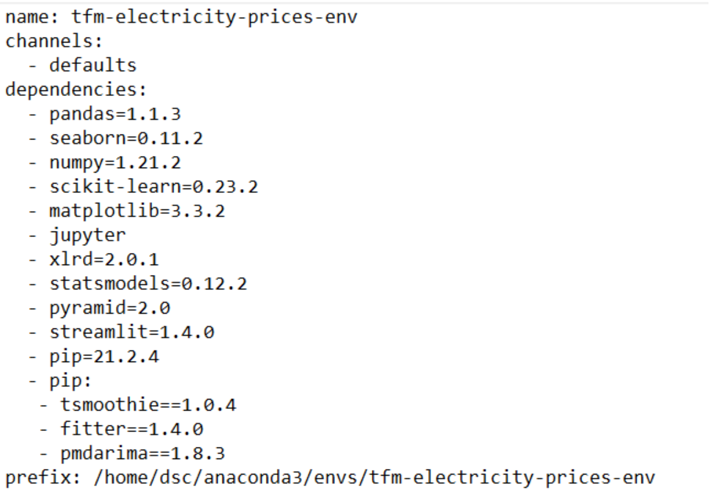
```
CarpetaCompartida/TFM/Data
└── [4.0K]  Precios de la electricidad
    ├── [4.0K]  Aemet2021-06-Por-Estaciones
    │   ├── [ 760]  Columnas-Unidades.txt
    │   ├── [ 32K]  DatosEstaciones-2021-06 [850 entries exceeds filelimit, not opening dir]
    │   └── [ 52K]  ListadoEstaciones.csv
    ├── [ 37M]  Aemet2021-06-Por-Estaciones.zip
    ├── [4.0K]  Aemet2021-10-Por-Estaciones
    │   ├── [ 760]  Columnas-Unidades.txt
    │   ├── [ 28K]  DatosEstaciones - 2021-10 [850 entries exceeds filelimit, not opening dir]
    │   └── [ 52K]  ListadoEstaciones.csv
    ├── [ 38M]  Aemet2021-10-Por-Estaciones.zip
    ├── [4.0K]  Archive
    │   └── [140K]  Datos_históricos_Futuros_petróleo_Brent.csv
    ├── [365K]  art_194_1 - Determinantes del precio de la electricidad en España.pdf
    ├── [1.7K]  Cifras oficiales de población resultantes de la revisión del Padrón municipal 2020.csv
    ├── [424K]  ComposicionPrecios.pdf
    ├── [146K]  Datos_históricos_Futuros_petróleo_Brent.csv
    ├── [4.0K]  Definitivo_2013 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivo_2014 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivo_2015 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivo_2016 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivo_2017 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivos_2018 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivos_2019 [26 entries exceeds filelimit, not opening dir]
    ├── [4.0K]  Definitivos_2020
    │   ├── [4.0K]  2020
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200101_20200131.xls
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200201_20200229.xls
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200301_20200331.xls
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200401_20200430.xls
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200501_20200531.xls
    │   │   ├── [ 16K]  PFMDIASM_TOD_20200601_20200630.xls
    │   │   ├── [ 17K]  PFMDIASM_TOD_20200701_20200731.xls
    │   │   ├── [ 17K]  PFMDIASM_TOD_20200801_20200831.xls
    │   │   ├── [ 17K]  PFMDIASM_TOD_20200901_20200930.xls
    │   │   ├── [ 17K]  PFMDIASM_TOD_20201001_20201031.xls
    │   │   ├── [ 17K]  PFMDIASM_TOD_20201101_20201130.xls
    │   │   └── [ 17K]  PFMDIASM_TOD_20201201_20201231.xls
```

Figure 3 - Directory Tree Structure

Methodology

1. Introduction: Conda Environment

Python 3.9 was the coding language used along this project. In order to maintain replicability, a Conda environment was created and saved in a .yml file. The installed packages and their version needed to run the code are presented below:

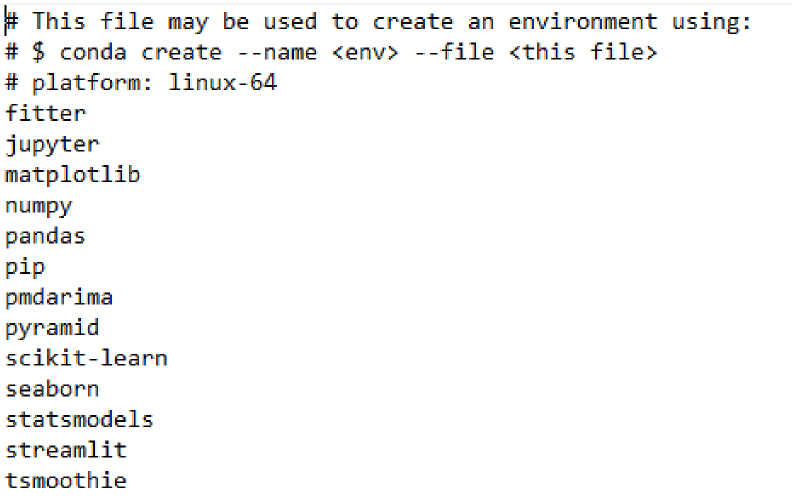


```
name: tfm-electricity-prices-env
channels:
  - defaults
dependencies:
  - pandas=1.1.3
  - seaborn=0.11.2
  - numpy=1.21.2
  - scikit-learn=0.23.2
  - matplotlib=3.3.2
  - jupyter
  - xlrd=2.0.1
  - statsmodels=0.12.2
  - pyramid=2.0
  - streamlit=1.4.0
  - pip=21.2.4
  - pip:
   - tsmoothie==1.0.4
   - fitter==1.4.0
   - pmdarima==1.8.3
prefix: /home/dsc/anaconda3/envs/tfm-electricity-prices-env
```

Figure 4 – Conda Environment "tfm-electricity-prices-env"

Additionally, some packages used were not available in Conda, so they were installed using Pip.

However, this .yml file "will not consistently produce environments that are reproducible across Mac OS, Windows, and Linux" (Pugh & Tocknell, 2019). For this reason, a requirements.txt file was also created, to ensure replicability across operating systems:



```
# This file may be used to create an environment using:
# $ conda create --name <env> --file <this file>
# platform: linux-64
fitter
jupyter
matplotlib
numpy
pandas
pip
pmdarima
pyramid
scikit-learn
seaborn
statsmodels
streamlit
tsmoothie
```

Figure 5 – Environment.txt

The environment can be installed using the following commands:

- using pip:

```
pip install -r requirements-tfm3.txt
```

- using Conda:

```
conda create --name tfm-electricity-prices-env --file requirements-tfm3.txt
```

2. The CRISP-DM Process Model

The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Shearer, 2000) was used in this project, with a constant interaction between all its phases.

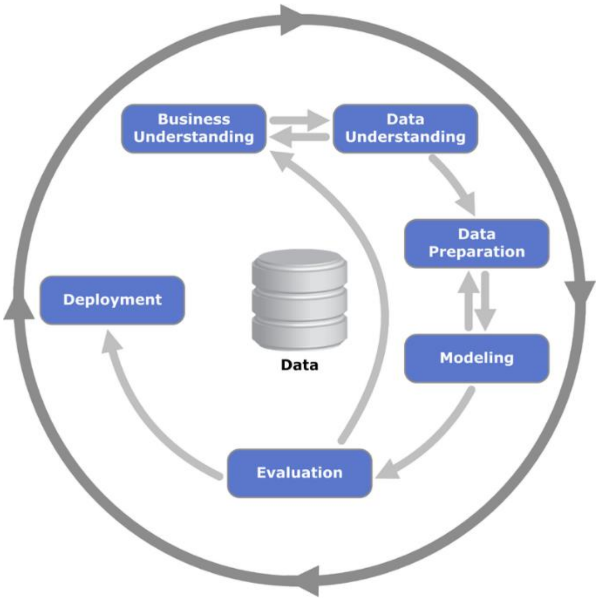


Figure 6 - Phases of CRISP-DM

2.1. Business Understanding and Data Understanding

The very first step was to understand how the electricity market works, and finding data that could be used to forecast the electricity price.

Firstly, data only up to 2020 was used, without finding or handling outliers. Then, data up to September 2021 was included, to be able to take into account the latest market changes and after that some outliers identification was done.

After all that, data was pre-processed in order to be used for modelling in the next step: data preparation.

Therefore, three scenarios can be identified in the Jupyter Notebooks for this project:

- Models and feature engineering using data from 2013 to 2020.
- Models and feature engineering using data from 2013 to 2021.
- Models and feature engineering using data from 2013 to 2021, identifying and handling outliers.

As it will be explained in the following sections, the second option listed above is the one with the best results, using the term "best" as "more accurate if compared to reality".

In the repository, there are also two notebooks called: "TEST data TMF Precio electricidad.ipynb" and "TMF Precio electricidad MVP.ipynb" that can be disregarded, as those were the first exploration and feature engineering with the data.

2.2. Data Preparation

Data preparation included all the steps between data gathering and modelling. This has been an iterative process and can be found at "TMF Precio electricidad - feature engineering and basic VIZ.ipynb" and "TMF Precio electricidad - feature engineering and basic VIZ - 2013-2021 - finding outliers.ipynb".

Firstly, all year's electricity prices, Brent crude oil future prices were loaded and merged into the same Data Frame using Pandas.

Secondly, the weather and population data were loaded, and this data required some transformations to get to a single daily value. As explained before, a median average by province was done along with a weighted average by each province population, so the values for those provinces with more inhabitants were more important.

It is important to mention here that only the peninsular weather stations were used, so the Balearic and Canary Islands, along with Ceuta and Melilla's weather stations were eliminated from the Data Frame.

When merging the weather data to the electricity and Brent data, all dates from January 2013 to the 4th of May of 2013 were removed, because those cannot be found at weather stations' data.

The reasoning behind all decisions made when pre-processing the data, are explained in the notebooks.

When the data was clean, constructed, integrated and formatted it was saved to a CSV file called "electricity_brent_weather1320.csv", so it could be used for modelling later.

In the second notebook, where data up to September 2021 was included, also outliers were detected and handled and three CSV files were created. One that included all integrated data from 2013 to 2021 called "electricity_brent_weather.csv", a second one called "electricity_price_no_outliers.csv" that only included the electricity prices without outliers and a last one called "electricity_brent_weather_no_outliers.csv", which is the same as the first one but the electricity prices here were transformed when they were an outlier.

The way outliers were identified and handled was that in cases where the absolute value between the price value and the mean of the series was above 3 times higher than the standard deviation of the series, those values were replaced with a value that is the median of the series.

This is a standard way of handling outliers, however, in this particular example, only some 2021 prices followed this rule because prices increased a lot from the beginning of 2021. Therefore, replacing this very high prices with the median of the series did not help the models to predict future prices that continued the increasing trend.

This is why, the predictions of the models that used data with outliers were more accurate if compared to reality.

2.3. Modelling

As mentioned above, three notebooks can be found in the repository with machine learning models for prediction, as well as statistical techniques to help understand the distribution of the data, test hypothesis, etc. Those are:

- "TFM Precio electricidad - Modelling.ipynb" with models using data up to 2020.
- "TFM Precio electricidad - Modelling - 2013-2021.ipynb" with models using data up to 2021 and not removing any outliers. This is the notebook whose results are shown in the front end.
- "TFM Precio electricidad - Modelling - 2013-2021 handling outliers.ipynb" with models using data up to 2021 and handling outliers.

The most important step here is using the date as the index of the Data Frame because time series techniques for forecasting were used.

Also, the division into train and test in the Data Frame cannot be done aleatory, but using the latest observations as the test set. Otherwise, time structure of the time series predictions would be ignored.

ARIMA models and Vector Autoregression Models (VAR) were used. In order to be able to use them, autocorrelation in the data was tested, as well and normality and stationarity in the time series.

Below there is an explanation of each model and some of the findings and steps performed. However, explanations on why decision were made are also included in the notebooks.

2.3.1. ARIMA

As explained by Hyndman and Athanasopoulos (2021), ARIMA models are a widely used approach to time series forecasting.

The key points when modelling using ARIMA with the electricity price time series are:

- Electricity prices time series using data up to 2020 (and up to 2021 removing outliers) are stationary and its "statistical properties do not depend on the time at which the series is observed” (Hyndman & Athanasopoulos, 2021). This means that when fitting ARIMA models, the “d” parameter or degree of first differencing involved would be 0.

- On the contrary, electricity prices time series using data up to September 2021 are non-stationary. This means that when fitting ARIMA models, the “d” parameter or degree of first differencing involved would be 1.
- Electricity prices do not follow a normal distribution, but the distribution is Gaussian-like in all scenarios, so it was assumed they are "normal". Some normalisation/standardisation techniques, like the Box-Cox transformation or the Min-Max scaler were applied to the electricity prices time series to try and make it “normal”, but without success.
- Electricity prices are not seasonal.
- Electricity prices time series using data up to 2020 (and up to 2021 removing outliers) dot not follow a particular trend, while electricity prices time series using data up to September 2021, seems to follow an increasing trend from 2019.

In each of the scenarios, 2 ARIMA models were fitted. All of them used a walk-forward validation. This methodology was explained by Brownlee (2019) in his article and basically means that the model is updated each time step that new data is available.

The first ARIMA model was fitted only choosing the degree of first differencing (depending on if the data was stationary or not), and using 10 for the order of the autoregressive part of the ARIMA and 0 for the order of the moving average part.

As written by Hyndman and Athanasopoulos (2021):

> The equation of the ARIMA model can be written as:
>
> $$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$
>
> *Figure 7 - ARIMA Equation*
>
> Where y'_t is the differenced series. This is an ARIMA model where
>
> $p =$ order of the autoregressive part;
>
> $d =$ degree of first differencing involved;
>
> $d =$ order of the moving average part (9.5 Non-seasonal ARIMA models)

In order to optimise the ARIMA models, then the function auto_arima() from the pdmarima package was used, returning they optimal p, d and d parameters. Afterwards, those parameters given by the auto_arima() function were used again the train another ARIMA model.

2.3.2. Vector Auto Regression (VAR)

Understanding the relationships between the final electricity price and the other variables, like temperatures, the price of the Brent crude oil futures, etc. was one of the objectives of this project.

Because of this, a VAR model was also trained with each of the scenario's data. This model assumes that the electricity price has a linear relationship with other time series (Hyndman & Athanasopoulos, 2021) and "are popular in economics and other sciences because they are flexible and simple models for multivariate time series data" (Lütkepohl, 2011).

This time, the time series are multivariate ($Y = T \times K$), where T denotes the number of observations and K the number of variables (Perktold, Seabold, & Taylor, 2022).

As stated by Perktold, Seabold, and Taylor (2022), the vector autoregression process estimates the relationships between the time series and their lagged values:

$$\begin{aligned} Y_t = \nu + A_1 Y_{t-1} + \ldots + A_p Y_{t-p} + u_t \\ u_t \sim \mathsf{Normal}(0, \Sigma_u) \end{aligned}$$

Figure 8 - VAR Equation

Where A_i is a $K \times K$ coefficient matrix.

The VAR class also assumes that all time series passed are stationary. This was not true for the electricity prices when using data up to 2021 and also was not the case for the Brent crude oil future prices. The Augmented Dickey-Fuller test was performed on all relevant time series to check stationarity.

To make all time series stationary and to remove seasonality from other variables, like weather variables, a Standard Scaler was used, followed by a first order differencing. The Standard Scaler standardised all features "by removing the mean and scaling to unit variance" (Scikit-learn developers, 2021).

The lag order selection for the VAR model was done using the AIC criteria.

After training the model and making predictions for all the features, the transformation made before to standardise them were inverted, so all variables were shown again in their original scale.

2.4. Model's Evaluation and Summary

The metrics used to evaluate all models were the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). These metrics tells us how accurate the predictions are and the amount of deviation from the actual values (Acharya, 2021).

In the repository there is a notebook called "Summary models metrics.ipynb", where a comparison of the metrics for all models was done.

Furthermore, in the repository there is also a folder called "models_pkl" where all the trained models are saved (while running the code in each modelling notebook these files are generated).

Summary of Main Results

All the predictions/forecasts can be found in the Modelling notebooks. Predictions were made for all the three scenarios explained before.

In the scenario where data used was only up to 2020, predictions showed better metrics but if we compare those with reality, they were not accurate.

Similar issue with the scenario where outliers were removed by using the median of the series, where predictions were not in line with reality. Furthermore, this scenario showed the worst metrics in the two ARIMA models used, but the best RMSE in the VAR model, compared to all 3 scenarios.

As explained before, the predictions made using data up to September 2021 without handling outliers is the scenario that was used for the front end. This is because predictions were more in line with reality in the short term than in the other two scenarios. However, the VAR model showed the worst RMSE for the Total price per MWh in €, compared to all scenarios.

Overall, models performed quite well and short-term predictions were somehow accurate. The issue with these kinds of models is that long-term forecasts tend to go to the mean of the data or will follow a straight line, depending on the parameters used to fit the models, as explained by Hyndman and Athanasopoulos (2021).

Conclusions

After deeply studying the data and using the previously mentioned models and machine learning techniques, one of the main conclusions drawn is that even if the right statistical methodologies and algorithms are used in Machine Learning, it is very difficult to predict or forecast, in this case future prices, when major changes in the market occur.

As all of the prediction/forecast machine learning techniques use historical data to learn about that data and then use historical patterns to predict, when the market goes through an unprecedented change, it is just not possible to predict the future with some kind of accuracy.

This can be seen in the predictions made by ARIMA and VAR models using data up to December 2020 and trying to predict 2021 prices. Those predictions are far away from reality because during 2021 we have seen prices never-before-seen.

Another point worth making, that impacted results, is all the decisions made through the project in terms of features used in the models. It is clear that some more variables could have been used to improve the performance, for example, the price of carbon emissions.

However, it is important to balance the amount of time needed to add more data from completely different sources into the datasets considering that, as previously mentioned, it was almost impossible to make accurate predictions for a very unpredictable market at the moment.

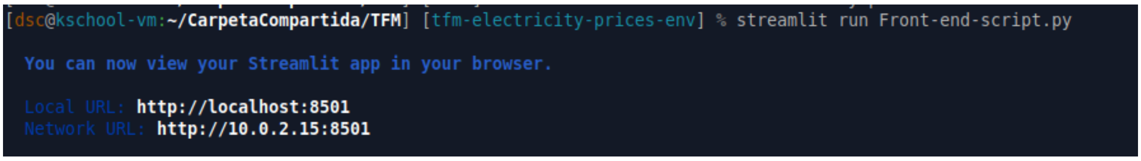
Thirdly, I would like to mention that in this particular case, handling or removing outliers (categorised as those prices whose absolute difference from the series mean is 3 times higher than the standard deviation) was not useful, because 2021 prices could be all categorised as outliers under this assumption. Therefore, removing them or changing their value would not help the models to predict accurately.

Nevertheless, I think this was a very interesting subject that was worth studying.

User Manual – Front End

An interactive app was done with Streamlit with the purpose of quickly showing the results of this project.

In order to run the app, you have to write the command "streamlit run Front-end-script.py" in the command line (while in the same directory where the python script is saved and after activating the conda environment):

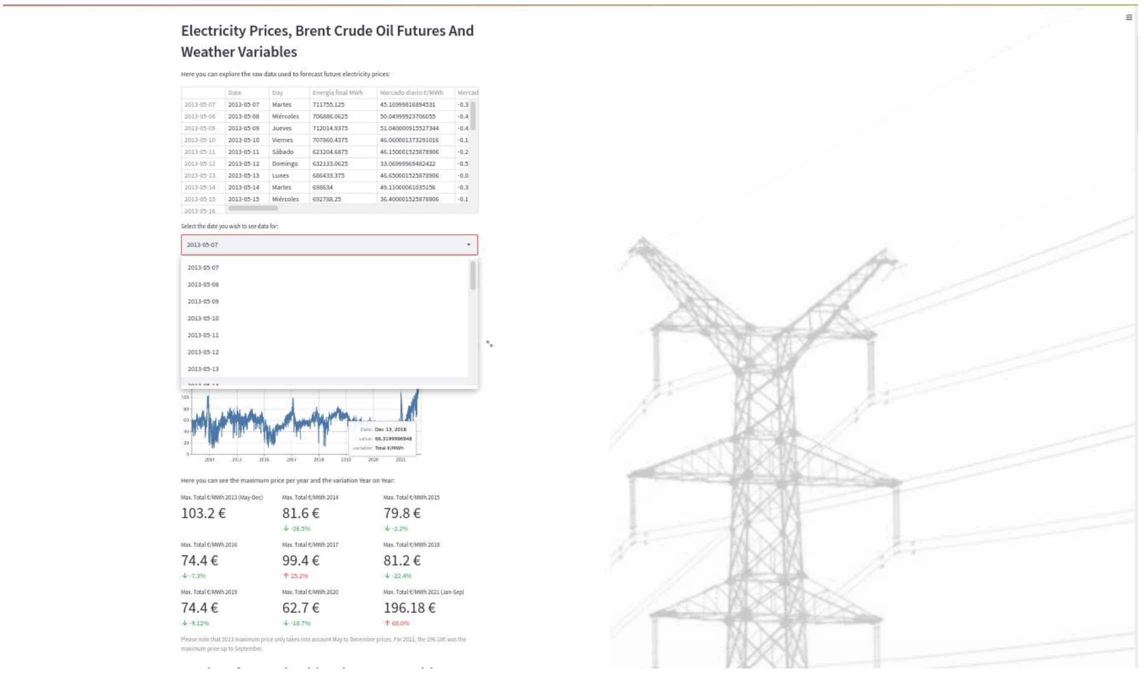


```
[dsc@kschool-vm:~/CarpetaCompartida/TFM] [tfm-electricity-prices-env] % streamlit run Front-end-script.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://10.0.2.15:8501
```

Figure 9 - Streamlit Run Command Line

This will open a browser tab with the app, which is structured in 4 sections.

First section is an introduction of the project, where the users can explore the raw data and look at an interactive plot with historical electricity prices. They can also filter the data frame by date to check a specific day and check what was the maximum electricity price of the year since 2013, along with the YOY variations:



Electricity Prices, Brent Crude Oil Futures And Weather Variables

Here you can explore the raw data used to forecast future electricity prices:

	Date	Day	Energía final MWh	Mercado diario €/MWh	Mercad
2013-05-07	2013-05-07	Martes	711755.125	45.16999816894531	-0.3
2013-05-08	2013-05-08	Miércoles	706886.0625	50.04999923706055	-0.4
2013-05-09	2013-05-09	Jueves	712014.9375	51.040000915527344	-0.4
2013-05-10	2013-05-10	Viernes	707860.4375	46.060001373291016	-0.1
2013-05-11	2013-05-11	Sábado	623204.6875	46.150001525878906	-0.2
2013-05-12	2013-05-12	Domingo	632133.0625	33.06999969482422	-0.5
2013-05-13	2013-05-13	Lunes	686433.375	46.650001525878906	-0.0
2013-05-14	2013-05-14	Martes	698634	49.11000061035156	-0.3
2013-05-15	2013-05-15	Miércoles	692788.25	36.400001525878906	-0.1
2013-05-16					

Select the date you wish to see data for: 2013-05-07

Here you can see the maximum price per year and the variation Year on Year:

Max. Total €/MWh 2013 (May-Dec)
103.2 €

Max. Total €/MWh 2014
81.6 €
↓ -26.5%

Max. Total €/MWh 2015
79.8 €
↓ -2.2%

Max. Total €/MWh 2016
74.4 €
↓ -7.3%

Max. Total €/MWh 2017
99.4 €
↑ 25.2%

Max. Total €/MWh 2018
81.2 €
↓ -22.4%

Max. Total €/MWh 2019
74.4 €
↓ -9.12%

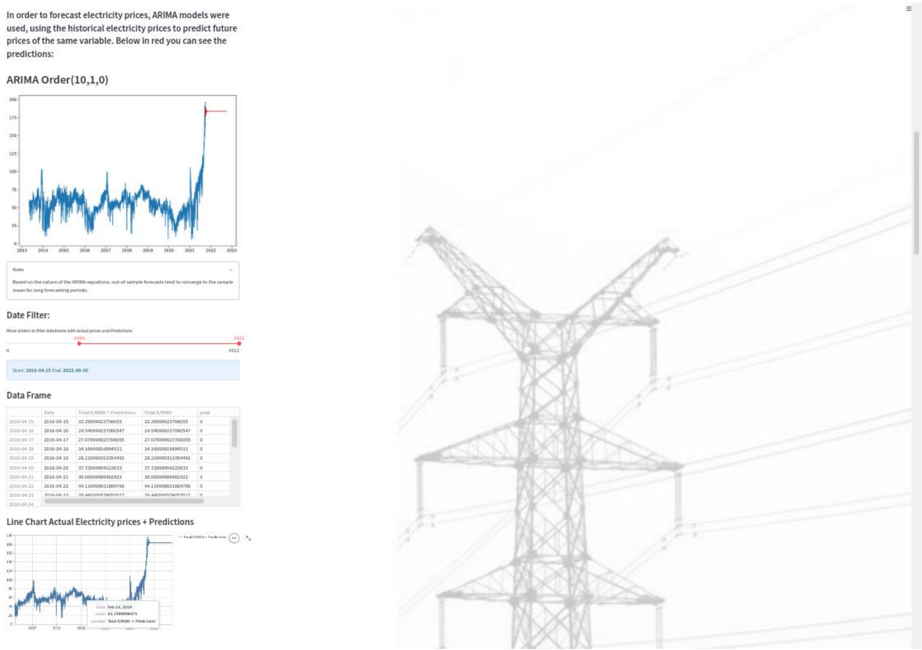
Max. Total €/MWh 2020
62.7 €
↓ -18.7%

Max. Total €/MWh 2021 (Jan-Sep)
196.18 €
↑ 68.0%

Please note that 2013 maximum price only takes into account May to December prices. For 2021, the 196.18€ was the maximum price up to September.

Figure 10 - Frontend First Section

If they scroll down, there is a brief explanation saying that ARIMA models were used to forecast electricity prices, and a plot of the historical prices along with the predictions in a different colour is shown. Below there is an interactive slider, where the users can choose the range of dates that they wish to see either historical prices or predictions. This slider filters a data frame and also a plot:



In order to forecast electricity prices, ARIMA models were used, using the historical electricity prices to predict future prices of the same variable. Below in red you can see the predictions:

ARIMA Order(10,1,0)

Notes

Based on the nature of the ARIMA equations, out-of-sample forecasts tend to converge to the sample mean for long forecasting periods.

Date Filter:

Move sliders to filter dataframe with Actual prices and Predictions

Start: 2016-04-15 End: 2022-09-30

Data Frame

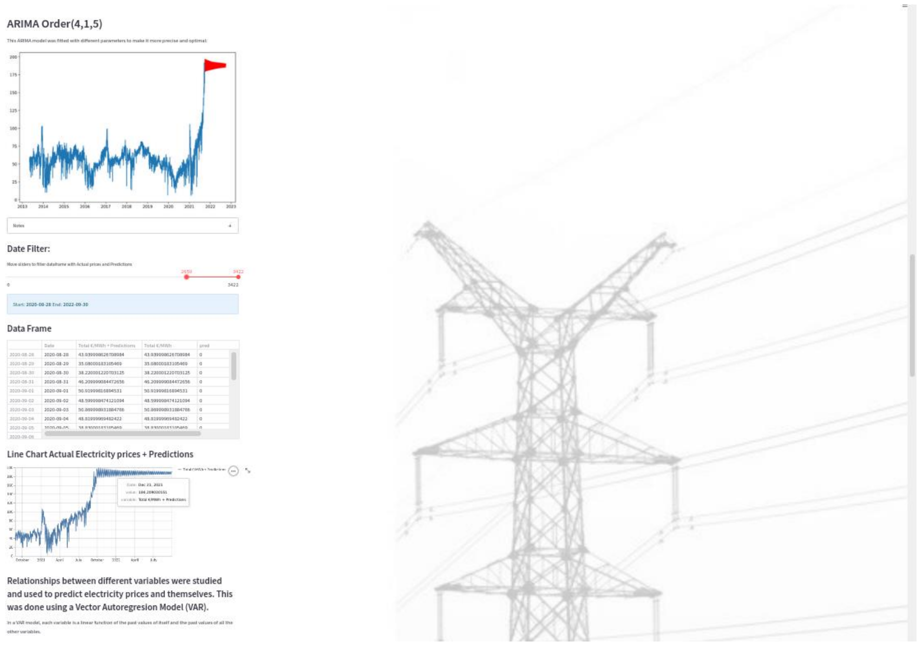
	Date	Total €/MWh + Predictions	Total €/MWh	pred
2016-04-15	2016-04-15	32.29999923706055	32.29999923706055	0
2016-04-16	2016-04-16	24.549999237060547	24.549999237060547	0
2016-04-17	2016-04-17	27.079999923706055	27.079999923706055	0
2016-04-18	2016-04-18	34.16999816894531	34.16999816894531	0
2016-04-19	2016-04-19	28.219999313354492	28.219999313354492	0
2016-04-20	2016-04-20	37.72999954223633	37.72999954223633	0
2016-04-21	2016-04-21	39.08999969482422	39.08999969482422	0
2016-04-22	2016-04-22	44.119998931884766	44.119998931884766	0
2016-04-23	2016-04-23	29.440000534057617	29.440000534057617	0
2016-04-24				

Line Chart Actual Electricity prices + Predictions

Date: Feb 14, 2019
value: 61.159999084
variable: Total €/MWh + Predictions

Figure 11 - Frontend Second Section

Following this, there is another section with the optimised ARIMA model, and the structure is the same as above:



ARIMA Order(4,1,5)

This ARIMA model was fitted with different parameters to make it more precise and optimal:

Notes +

Date Filter:

Move sliders to filter dataframe with Actual prices and Predictions

Start: 2020-08-28 End: 2022-09-30

Data Frame

	Date	Total €/MWh + Predictions	Total €/MWh	pred
2020-08-28	2020-08-28	43.939998626708984	43.939998626708984	0
2020-08-29	2020-08-29	35.08000183105469	35.08000183105469	0
2020-08-30	2020-08-30	38.220001220703125	38.220001220703125	0
2020-08-31	2020-08-31	46.209999084472656	46.209999084472656	0
2020-09-01	2020-09-01	50.91999816894531	50.91999816894531	0
2020-09-02	2020-09-02	48.599998474121094	48.599998474121094	0
2020-09-03	2020-09-03	50.869998931884766	50.869998931884766	0
2020-09-04	2020-09-04	48.81999969482422	48.81999969482422	0
2020-09-05	2020-09-05	38.83000183105469	38.83000183105469	0
2020-09-06				

Line Chart Actual Electricity prices + Predictions

Date: Dec 21, 2021
value: 184.209030151
variable: Total €/MWh + Predictions

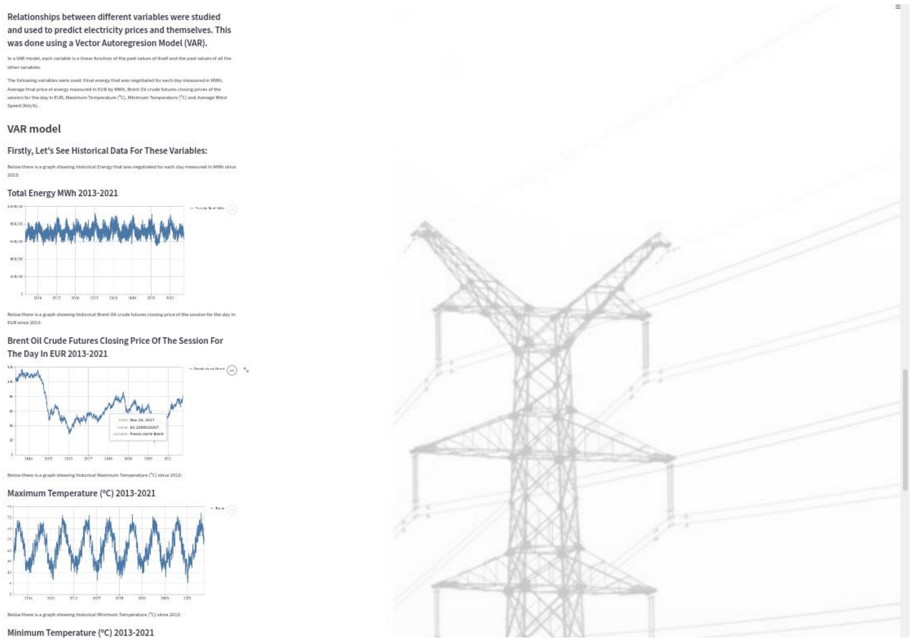
Relationships between different variables were studied and used to predict electricity prices and themselves. This was done using a Vector Autoregresion Model (VAR).

In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables.

Figure 12 - Frontend Third Section

Lastly, a brief explanation of how electricity prices were forecasted alongside with some other relevant variables are shown and users can firstly, check in interactive plots historical values for all those variables and, secondly, use a double ended slider to filter a data frame with the predictions for all those relevant variables.

This slider also filters a plot with the predictions for the electricity prices.



Relationships between different variables were studied and used to predict electricity prices and themselves. This was done using a Vector Autoregresion Model (VAR).

In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables.

The following variables were used: Final energy that was negotiated for each day measured in MWh, Average final price of energy measured in EUR by MWh, Brent Oil crude futures closing prices of the session for the day in EUR, Maximum Temperature (°C), Minimum Temperature (°C) and Average Wind Speed (Km/h).

VAR model

Firstly, Let's See Historical Data For These Variables:

Below there is a graph showing historical Energy that was negotiated for each day measured in MWh since 2013:

Total Energy MWh 2013-2021

Below there is a graph showing historical Brent Oil crude futures closing price of the session for the day in EUR since 2013:

Brent Oil Crude Futures Closing Price Of The Session For The Day In EUR 2013-2021

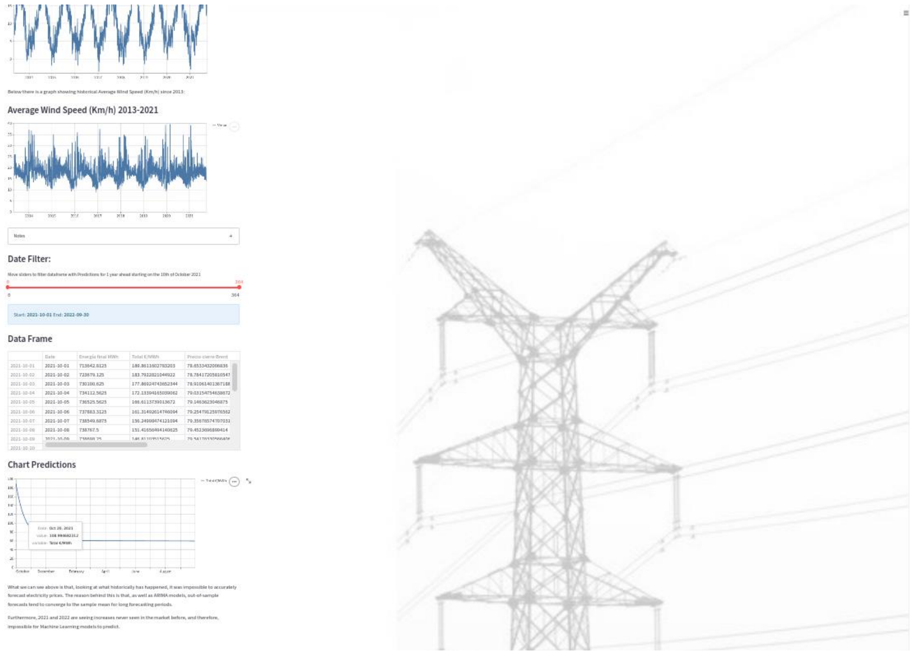
Below there is a graph showing historical Maximum Temperature (°C) since 2013:

Maximum Temperature (ºC) 2013-2021

Below there is a graph showing historical Minimum Temperature (°C) since 2013:

Minimum Temperature (ºC) 2013-2021

Figure 13 - Frontend Fourth Section 1



Below there is a graph showing historical Average Wind Speed (Km/h) since 2013:

Average Wind Speed (Km/h) 2013-2021

Date Filter:

Move sliders to filter dataframe with Predictions for 1 year ahead starting on the 10th of October 2021

Start: 2021-10-01 End: 2022-09-30

Data Frame

	Date	Energía final MWh	Total €/MWh	Precio cierre Brent
2021-10-01	2021-10-01	713642.8125	189.8611602783203	78.6533432006836
2021-10-02	2021-10-02	723679.125	183.7922821044922	78.78417205810547
2021-10-03	2021-10-03	730100.625	177.86924743652344	78.91061401367188
2021-10-04	2021-10-04	734112.5625	172.13394165039062	79.03154754638672
2021-10-05	2021-10-05	736525.5625	166.6113739013672	79.1463623046875
2021-10-06	2021-10-06	737883.3125	161.31492614746094	79.25479125976562
2021-10-07	2021-10-07	738549.6875	156.24998474121094	79.35676574707031
2021-10-08	2021-10-08	738767.5	151.41656494140625	79.4523696899414
2021-10-09	2021-10-09	738698.25	[illegible]	[illegible]
2021-10-10				

Chart Predictions

What we can see above is that, looking at what historically has happened, it was impossible to accurately forecast electricity prices. The reason behind this is that, as well as ARIMA models, out-of-sample forecasts tend to converge to the sample mean for long forecasting periods.

Furthermore, 2021 and 2022 are seeing increases never seen in the market before, and therefore, impossible for Machine Learning models to predict.

Figure 14 - Frontend Fourth Section 2

At the bottom there is a brief conclusion of the project that the users can read.

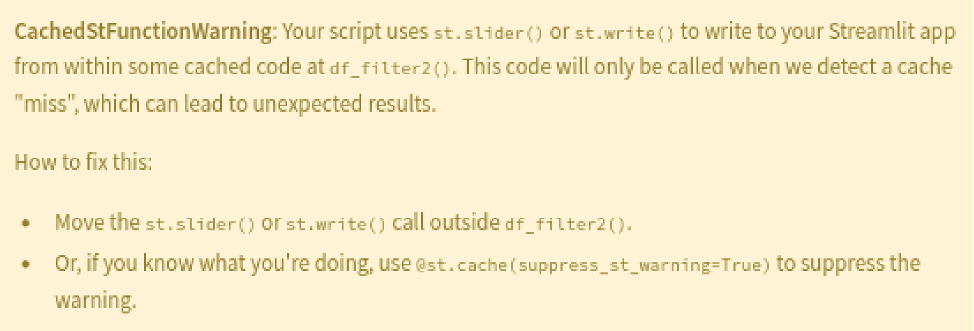
The main target of this app is anyone of the public who is affected by the progression of electricity prices and is interested in the possibility of forecasting them.

Because of this, my app was also deployed in Streamlit Cloud and anyone with the following URL can access it: https://share.streamlit.io/paulacervilla/streamlit-app/Front-end-script.py

To be able to do this, I created a separated GitHub repository, only with the script and the libraries needed to run it in a txt file. This repository can be found here: https://github.com/PaulaCervilla/streamlit-app

In the main repository, there is also a folder called "front end streamlit" where all the CSVs with data needed to run the app are saved.

It is important to mention that the app takes some time when you use the sliders. This is because the cache functionality of Streamlit could not be used because the sliders were defined within a function:



CachedStFunctionWarning: Your script uses `st.slider()` or `st.write()` to write to your Streamlit app from within some cached code at `df_filter2()`. This code will only be called when we detect a cache "miss", which can lead to unexpected results.

How to fix this:

- Move the `st.slider()` or `st.write()` call outside `df_filter2()`.
- Or, if you know what you're doing, use `@st.cache(suppress_st_warning=True)` to suppress the warning.

Figure 15 - Streamlit Cache Warning

Bibliography

Acharya, S. (2021). *What are RMSE and MAE?* Retrieved from Towards Data Science: https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383

Brownlee, J. (2019). *How To Backtest Machine Learning Models for Time Series Forecasting*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/

Comisión Nacional de los Mercados y la Competencia. (2021). *Electricity market*. Retrieved from https://www.cnmc.es/ambitos-de-actuacion/energia/mercado-electrico

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). Retrieved from https://otexts.com/fpp3/

Investing.com. (2021). *Datos históricos Futuros petróleo Brent*. Retrieved from https://es.investing.com/commodities/brent-oil-historical-data

Lütkepohl, H. (2011). Vector Autoregressive Models. *Lovric M. (eds) International Encyclopedia of Statistical Science.* doi:https://doi.org/10.1007/978-3-642-04898-2_609

OMIE. (2021). *About us*. Retrieved from https://www.omie.es/en/sobre-nosotros

OMIE. (2021). *Electricity market*. Retrieved from https://www.omie.es/en/mercado-de-electricidad

Perktold, J., Seabold, S., & Taylor, J. (2022). *Vector Autoregressions tsa.vector_ar*. Retrieved from statsmodels v0.14.0.dev0 (+285): https://www.statsmodels.org/dev/vector_ar.html

Pugh, D., & Tocknell, J. (2019). *Introduction to Conda for (Data) Scientists*. Retrieved from https://github.com/kaust-vislab/introduction-to-conda-for-data-scientists

Scikit-learn developers. (2021). *sklearn.preprocessing.StandardScaler*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5,* 13-22. Retrieved from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf