UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

240209 - CIÈNCIA DE DADES APLICADA A SISTEMES ELÈCTRICS

Report

Qiang Deng

Juan Sebastián Moreno Ordóñez

Sara Kiprijanova

December 2022

Introduction to the project work

The purpose of this project is to forecast the day-ahead electricity price in Spain by creating a supervised machine learning (ML) model. The dataset used to build and train the ML model was obtained from Red Espana[1], and it includes:

- Hourly consumption in MW – Spain 2019
- Hourly generation by type – Spain 2019
- Hourly generation in MW – Spain 2019

The datasets obtained contain assorted data in different categories which for the purpose of the project, were analyzed, filtered and processed before being used as the input data for the ML model, while the desired output value was the electricity price in EUR/MWh.

To build the required ML model, the following steps were obtained:

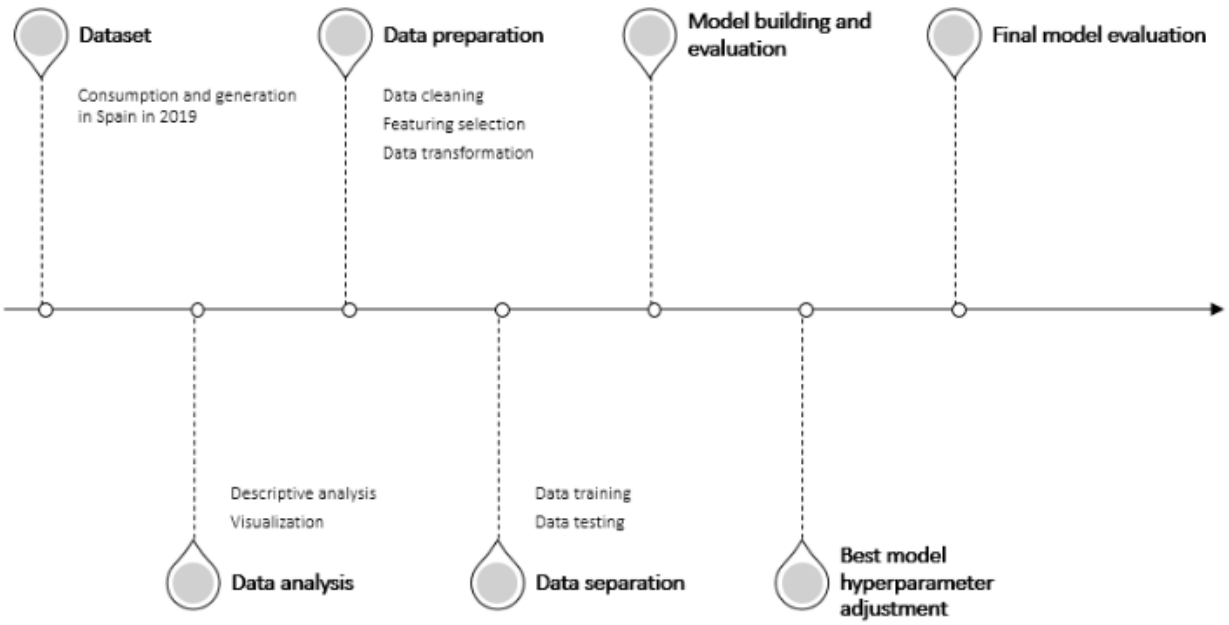

Figure 1. ML modelling process.

[1] Spanish conglomerate of companies that acts as the operator of the electrical system within the electrical energy market. Data access: https://www.esios.ree.es/es/mercados-y-precios

The ML model was built in Python and a separate Jupyter notebook containing the code is submitted in addition to this report. In the next part, this report presents each of the steps taken in detail and explains the actions and decisions taken.

Building the model

The European electricity price market has become incredibly volatile due to several reasons among which source of generation and ease of storage (Tschora 2022), and even more now due to global occasions such as the COVID19 pandemic and the consequences of the gas cutout due to the war between Russia and Ukraine (Liboreiro and de Fillippis 2021). Therefore, the ability to predict the electricity price on the market is of huge importance.

Building a ML model to predict energy prices requires a set of attributes that appear uncommon but have proven to be useful, such as: balance between production and consumption, dependence of the consumption on the time (hour of the day, day of the week, time of the year), load and generation that are influenced by external weather conditions (especially for penetration of renewables) and influence of neighboring markets (Garcia et al 2018).

We used predictive model examples from 2 papers: Tschora et al 2022 and Garcia et al 2018, to test already applied models for the same purpose and evaluate whether the same are compatible with the task assigned and the data set we have found. The explanation and results are provided in the next part of the report.

Working with the chosen dataset

The chosen dataset has been retrieved from the above-mentioned source. The building of our data frame has taken the following shape:

As Index values the date-time have been selected, these being hourly values for the whole of 2019.

Columns are composed of the following data:

- Hourly energy prices in Spain in 2019 [EUR/MWh].
- Nuclear generation in Spain in 2019 [MW].
- Hourly consumption in Spain in 2019 [MW].
- Solar power generation in Spain in 2019 [MW].
- Wind power generation in Spain in 2019 [MW].
- Energetic balance with France in 2019 [MW].
- Energetic balance with Portugal in 2019 [MW].
- Biogas-power generation in Spain in 2019 [MW].
- Biomass-power generation in Spain in 2019 [MW].
- Coal-power generation in Spain in 2019 [MW].
- CCP generation in Spain in 2019 [MW].
- Total exchange balance [MW].
- Fossil oil-energy generation in Spain in 2019 [MW].
- Gas prices in Spain 2019 from MIBGAS web (more recently added).

Data analysis

We have a dataset containing the electricity price in Spain in 2019, which is divided hourly from 1/1/2019 - 31/12/2019.

The purpose of this section is to analyze the trend of price changes throughout the year, to be able to understand what kind of trend to expect in the further forecasting, and to have valid data to compare our results to. Therefore, the following observations are obtained:

1. Time-series decomposition: to understand the series trend in price change over the year.
2. Monthly average price - to understand when the price of electricity was highest and lowest.
3. Generation power mix in the highest and lowest-price months - to understand what source generates expensive and cheap energy.

First, we want to look at whether the monthly electricity price displays seasonality and a trend - therefore we use the seasonal_decompose() function. This function breaks down a time series into 3 components: trend, seasonality, and random noise.

The visualization is presented in Figure 2:

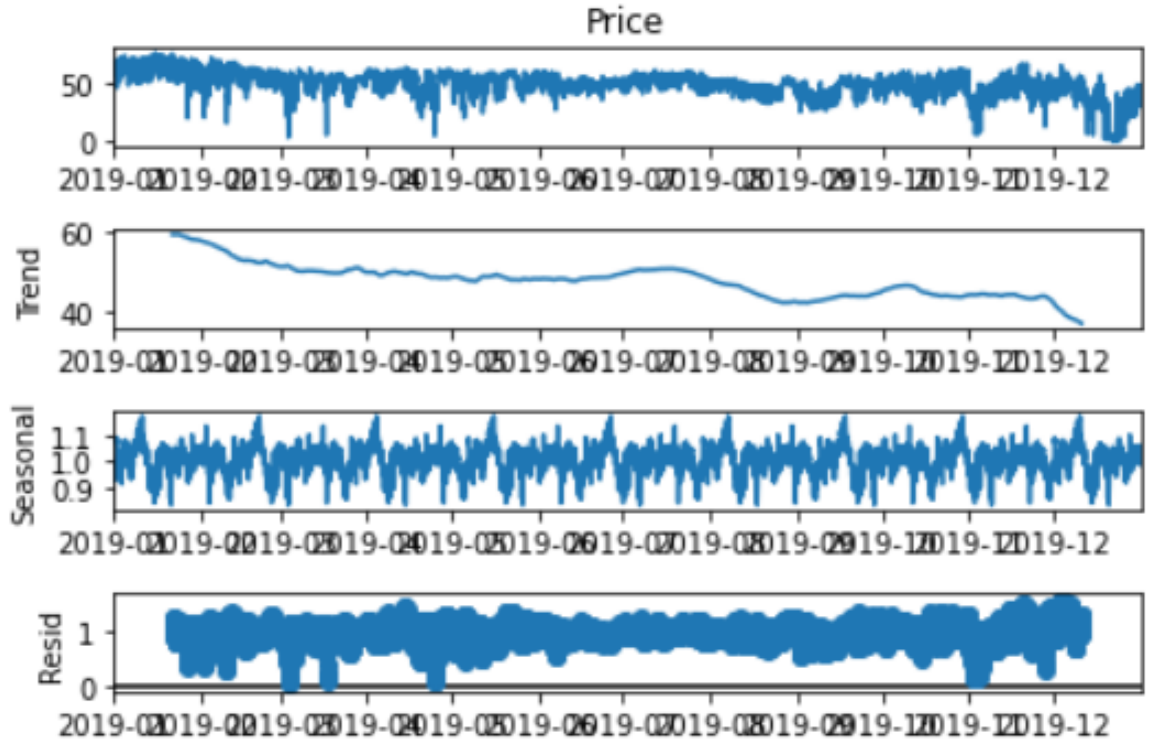

Figure 2. Time-series decomposition of the electricity price in Spain in 2019

From the results obtained, looking at the Trend component, it can be concluded that the price tends to decrease as the year progresses. However, there is a pattern in price high peaks approximately every 45 days and low peaks a few times a month, which shows that the price fluctuates a lot daily.

An information that we can work with is the trend for the price to decrease as the year progresses, and it is necessary to figure out what influences this change. Therefore, in the next part, the average price per month has been obtained, as well as the cheapest and most expensive months have been presented.

```
#Averge electricity price per month
months=df_hourprice.index.month
monthlyAvg=df_hourprice.groupby(months).Price.mean()
print(monthlyAvg)
```

```
datetime
1     61.959852
2     54.020491
3     48.836116
4     50.403611
5     48.393992
6     47.161236
7     51.464301
8     44.951815
9     42.130111
10    47.152325
11    42.187181
12    33.842970
Name: Price, dtype: float64
```

It can be observed that the trend has shown an extreme case where the average price of electricity in Spain in December is as twice as cheap as it has been at the beginning of the year in January. This can be because of several reasons, and the first one we are going to explore is how the generation mix in these two months has contributed.

A new dataset containing the generation by source has been created:

datetime	hydro	nuclear	oil	solar	wind
2018-12-31 23:00:00+00:00	23459.000000	23459.000000	23459.000000	23459.000000	23459.000000
2019-01-01 00:00:00+00:00	22781.000000	22781.000000	22781.000000	22781.000000	22781.000000
2019-01-01 01:00:00+00:00	21448.500000	21448.500000	21448.500000	21448.500000	21448.500000
2019-01-01 02:00:00+00:00	20262.166667	20262.166667	20262.166667	20262.166667	20262.166667
2019-01-01 03:00:00+00:00	19463.500000	19463.500000	19463.500000	19463.500000	19463.500000

To visualize the generation by source, the data has been plotted and the generation during January and December has been observed.

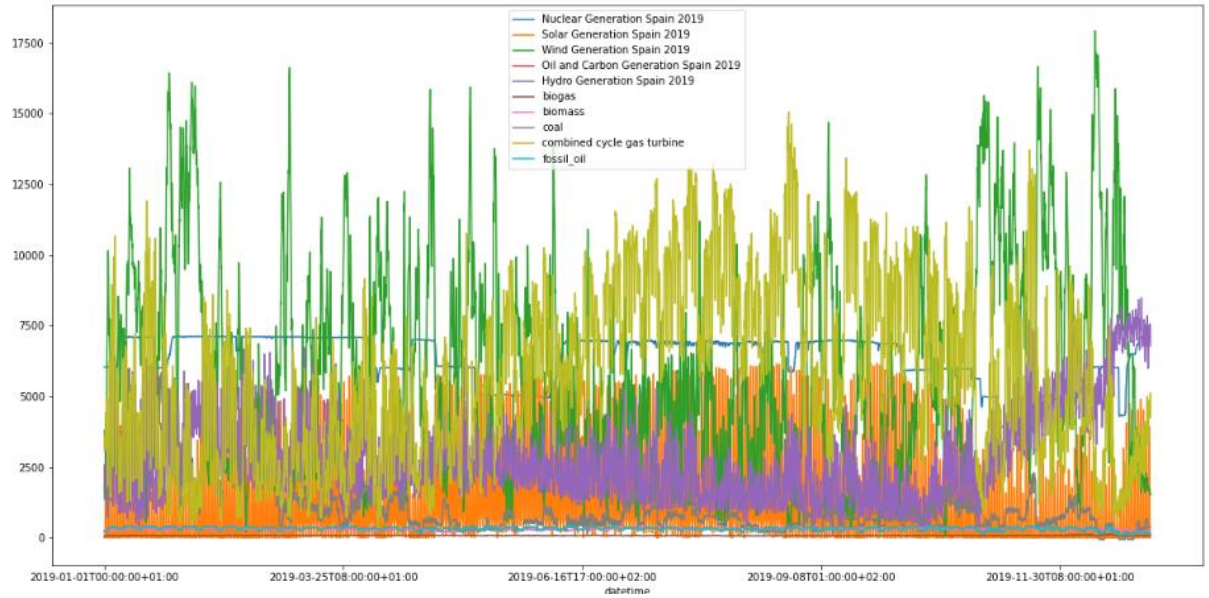

Figure 3. Hourly generation by source in Spain in 2019.

It can be noted that most of the generation in Spain during the winter months came from wind or combined cycle power plants, therefore these two sources were closely observed:

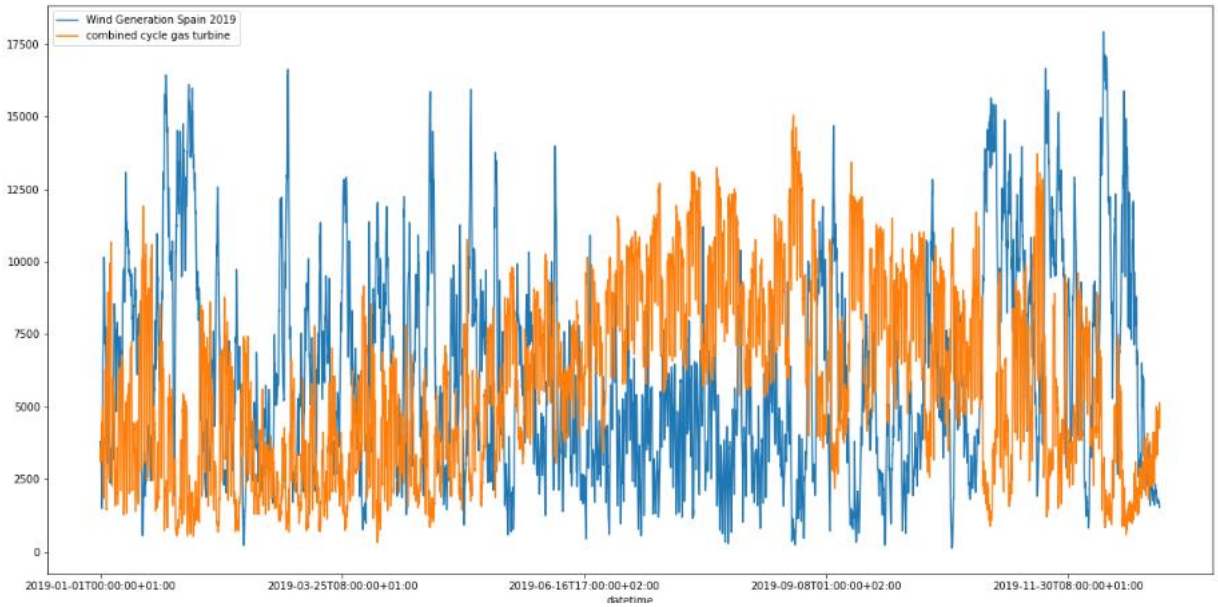

Figure 4. Hourly generation by wind and combined cycle power plants in Spain in 2019.

According to the obtained results, the energy during January was generated by wind and combined cycle sharing almost an equal amount, while in December it was mostly generated by wind sources, therefore it can be concluded that the wind generation conditions contributed to lower price of the energy towards the end of the year.

A Quick look into our database gives a general overview of the kind of data we are analyzing. The density of each selected parameter can be observed in the following figure:

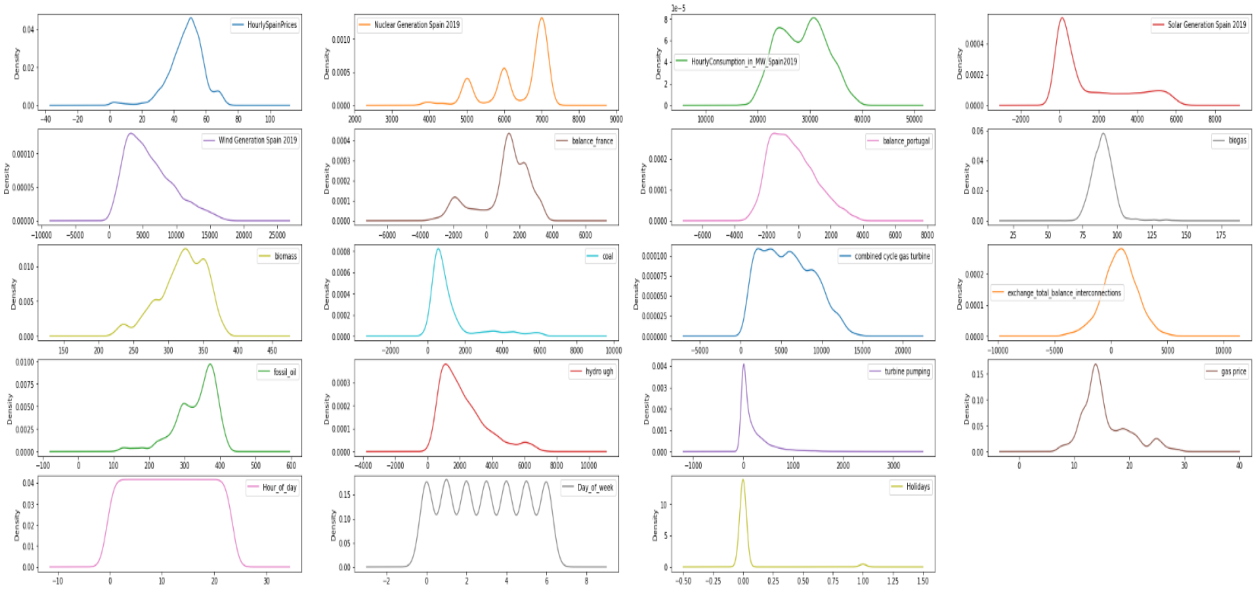

Figure 5. Density of selected parameters from the data analyzed.

Data preparation

From our data set we encountered some energy generation profiles containing two different indexes. Some data was double quantified since total demand and generation were classified in the same original data frame for each energy type generation. After data cleaning we have assumed for demand just the overall national value and each production type has been separated and assigned to its own column:

```
(8760, 16)
HourlySpainPrices                          float64
Nuclear Generation Spain 2019              float64
HourlyConsumption_in_MW_Spain2019          float64
Solar Generation Spain 2019                float64
Wind Generation Spain 2019                 float64
balance_france                             float64
balance_portugal                           float64
biogas                                     float64
biomass                                    float64
coal                                       float64
combined cycle gas turbine                 float64
exchange_total_balance_interconnections    float64
fossil_oil                                 float64
hydro ugh                                  float64
turbine pumping                            float64
gas price                                  float64
dtype: object
```

Figure 6. Final data frame columns.

The Pearson correlation method is then able to provide us with the following correlation heat map:

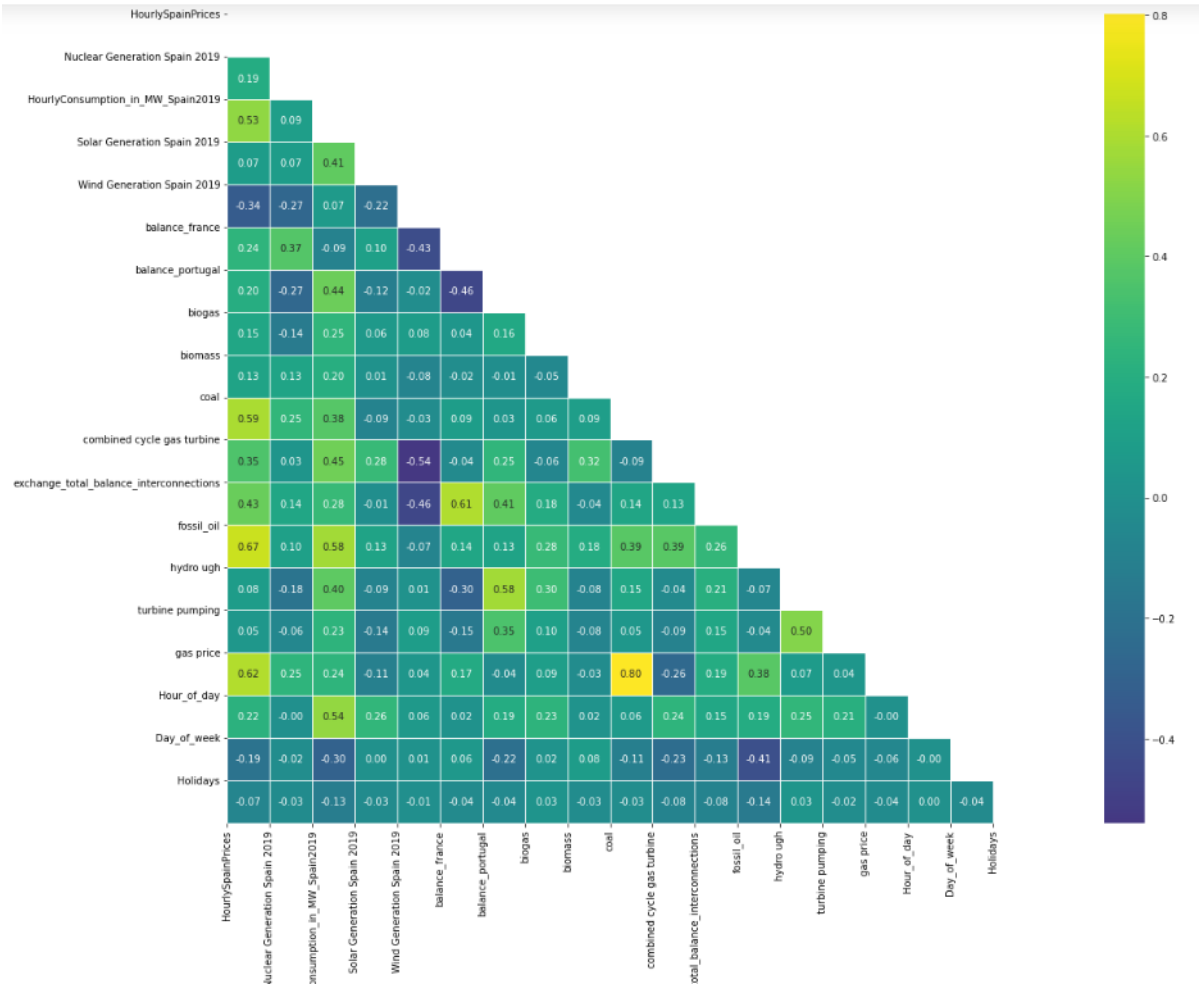

Figure 7. Pearson correlation method

We have assumed a standard correlation classification for the following values:

- Correlations with coefficients higher than 0.5 indicate a strong correlation.
- Correlations with coefficients between 0.3 and 0.49 indicate a medium correlation.
- Correlations with coefficients lower than 0.29 represent a small correlation.

Some of the obtained correlations were expected like for example a high increase in wind production reduces strongly the generation of CCP gas turbines. Also, when looking directly at the most relevant correlations to the prices in Spain, we find coal, fossil oil, the interconnections with France and Portugal and the hourly time frame to be more present. Additionally, we added the gas prices in 2019 after our first simulations to make our prediction more exact. This is reflected in the correlation matrix as well, where the price of gas is a staring variable. Interestingly, the days of the week and holidays do not pose a strong relationship factor according to the heat map. This shows that periodicity may play a role in forecasting, however not a direct one.

Data separation

Data separation will be conducted after the analysis of the correlation heatmap. This will alter the results depending on which variables are dropped from the data frame. Ideally the variation will not be strong, and it will decrease the amount of data needed for our model.

Time-Series analysis

In our prediction model, we have implemented an autocorrelation function with a lag frame of 48 hours. This one works along a partial autocorrelation function of a 24-hour lag.

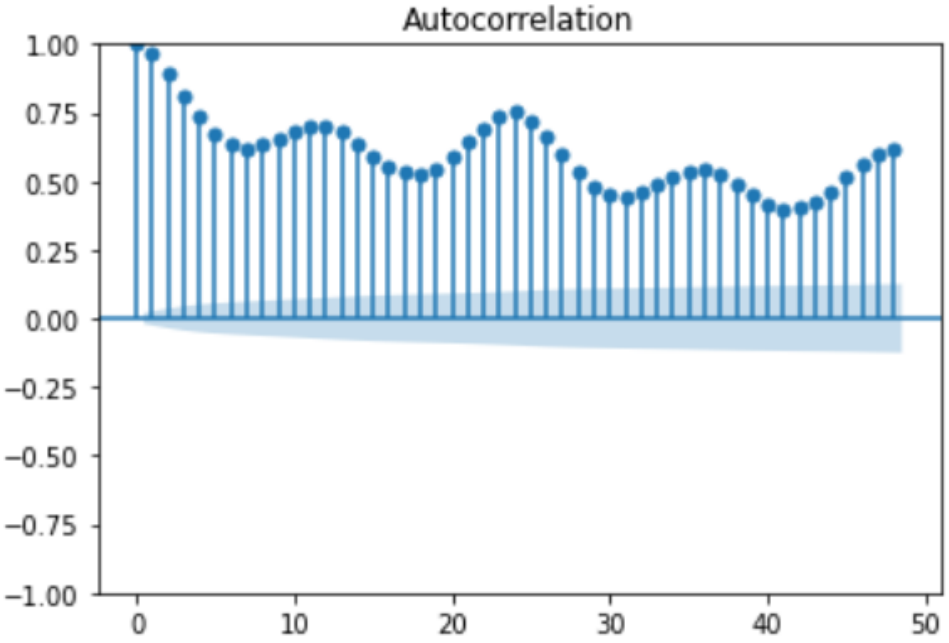

Figure 8. Autocorrelation function with a 48 hours' time lag.

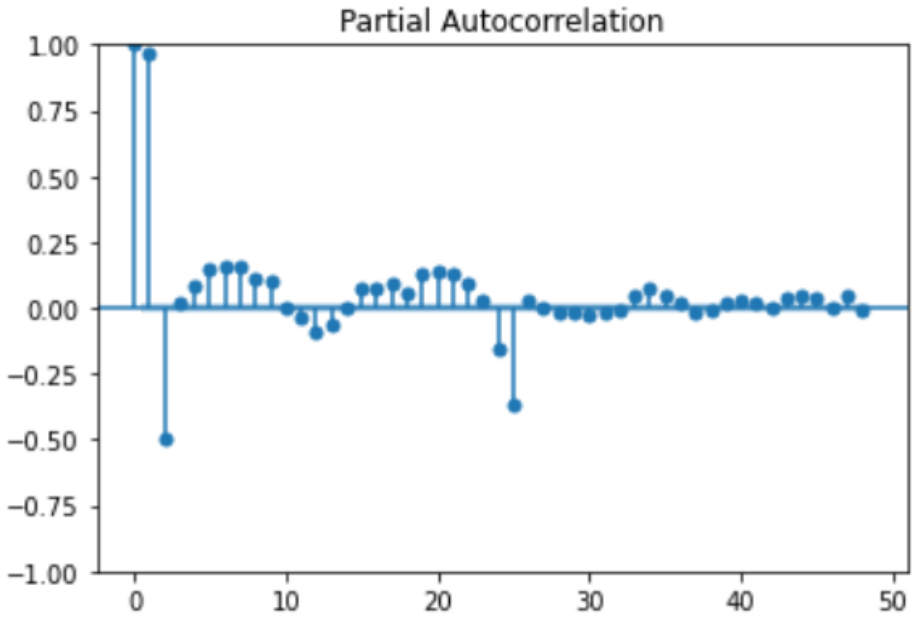

Figure 9. Partial autocorrelation function with a 48 hours' time lag.

Based on this we have added a lag of 48 hours and one week to the relevant energy source variables from the autocorrelation map (coal and fossil).

Model building and evaluation

For building the model we first select the target value and the attributes that we will use for the prediction:

```
X = dataset.drop(['HourlySpainPrices'], axis=1) # Attributes
y = dataset['HourlySpainPrices'] # Targets
```

We then normalize all the parameters to one with help of the sklearn preprocessing:

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X_df = X.copy()
X_scaled = pd.DataFrame(scaler.fit_transform(X_df))
X_scaled.columns = X_df.columns
X_scaled
```

Having then X_scaled as our attributes in ranges from 0 until 1. This is similarly done for our target variable. Then the model is built, and the data is divided into training, validation, and test data.

We obtained 209895 rows for the Train, 840 rows for the Test and 89985 rows for the Validation.

And the following models are loaded:

- MLP regression.
- Random Forest regressor.
- SVR.
- XGB regressor.
- K Neighbor.
- GBR.

The results are displayed as following for the r2 algorithm and root mean squared error (RMSE):

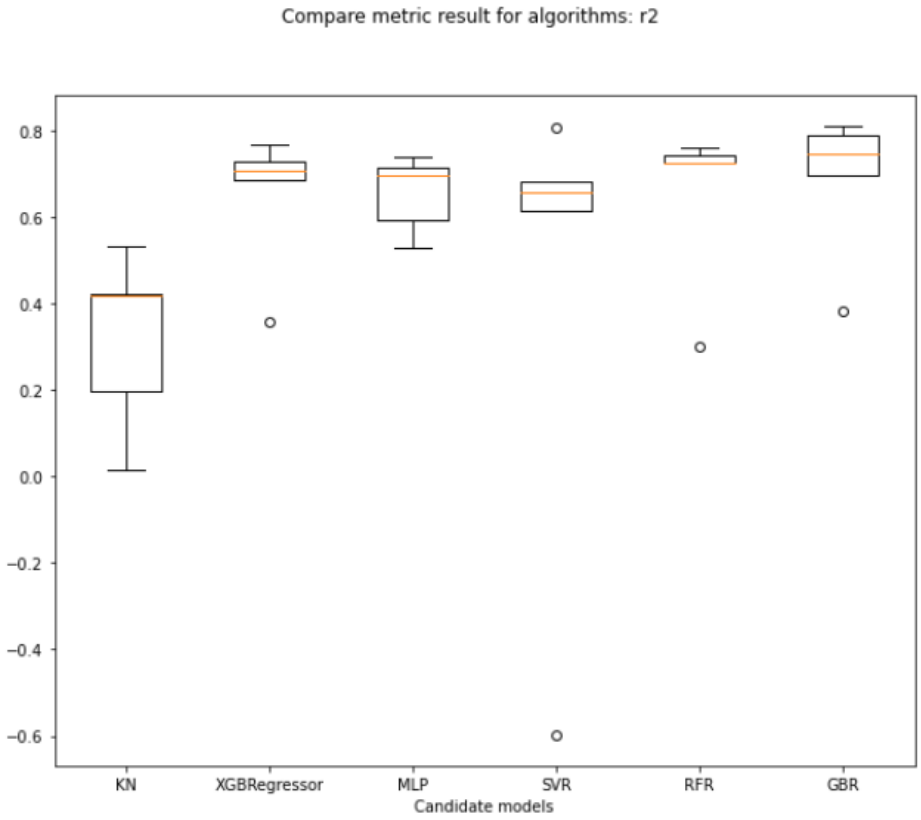

Figure 10. Comparison of metric results for the r2 algorithm.

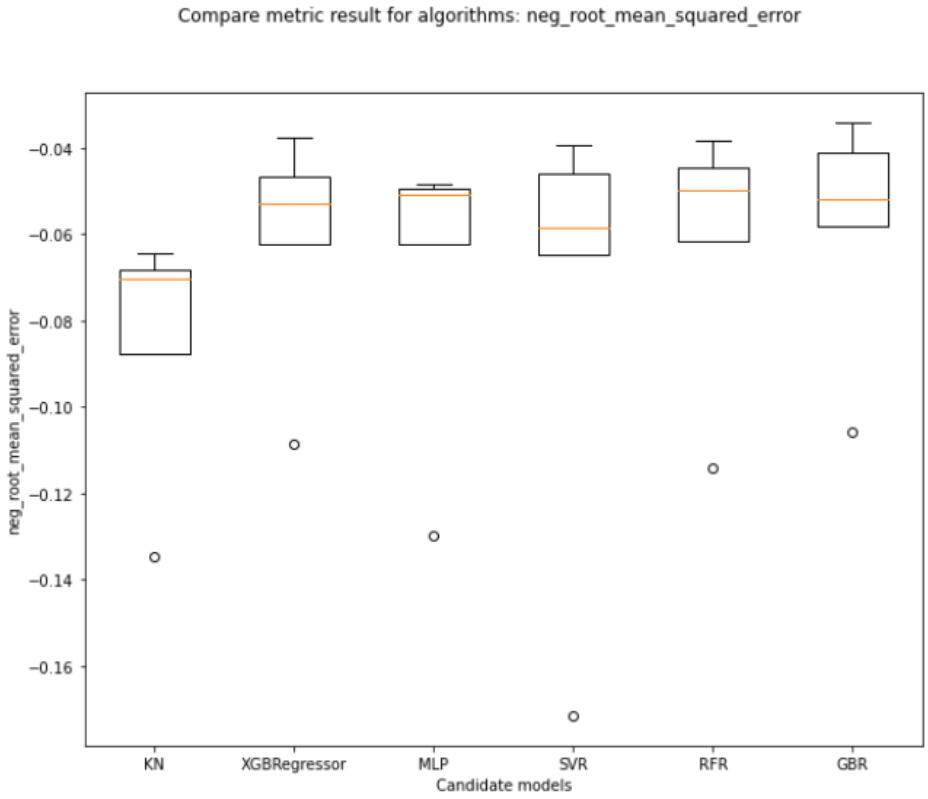

Figure 11. Comparison of metric results for the RMSE algorithm.

The r2 metric shows us how well the target value can be predicted in a normalized scala. Here we see a similarity between all models except KN. However, the density distributions are different in each model, making it not too trivial to choose the best model based on the boxplot. Also, it is noticeable that the GBR model has a slightly higher score for the median, this could be interesting to test the model on a simulation and see the results, but we must proceed with caution because finetuning can be a game changer when having so many similar results for different models.

The RMSE metric tells us about the error from each model and the prediction accuracy in general terms. Looking into it, we observe a smaller error for RFR. However, symmetry could be an issue and considering that we are having results in ranges smaller than 0.1, we have decided to test different predictions and choose the one that provides the smallest error at the end.

Best model hypermeter adjustment

For choosing our final model, we try different ones and then perform fine tuning for the hyperparameters. This is done through cross-validation. We give a range of tests for each parameter of the model, then they are tested and compared through a scoring mechanism. Finally, the results for the best model are printed. One example can be seen as follows with our first prediction model GBR:

```
from scipy.stats import uniform as sp_randFloat
from scipy.stats import randint as sp_randInt
model_1 = GradientBoostingRegressor()
scoring='r2'
parameters = {'learning_rate': sp_randFloat(),
                  'subsample'    : sp_randFloat(),
                  'n_estimators' : sp_randInt(100, 1000),
                  'max_depth'    : sp_randInt(4, 10)
                 }


cross_validation = KFold(n_splits=5, shuffle=False)
my_cv = cross_validation.split(X_val)
Rsearch = RandomizedSearchCV(estimator=model_1, param_distributions = parameters ,scoring=scoring, cv=my_cv)
Rsearch.fit(X_val, y_val)

print("Best result: %f using the following hyperparameters %s" % (Rsearch.best_score_, Rsearch.best_params_))
means = Rsearch.cv_results_['mean_test_score']
stds = Rsearch.cv_results_['std_test_score']
params = Rsearch.cv_results_['params']
```

```
Best result: 0.664242 using the following hyperparameters {'learning_rate': 0.16571203110911847, 'max_depth': 8, 'n_est
s': 429, 'subsample': 0.8628568860483482}
```

Figure 12. Final model testing between candidates and hyperparameter tuning.

These results are adopted for the model and the evaluation is run again.

We were able to obtain the following errors for each prediction:

Model/Error	GBR	SVR	AdaBoost + SVR
RMSE	10.59	3.53	2.28
MAE	10.11	2.04	1.96
MAPE	28.03	7.95	5.15

Since it reaches the lowest MAE from the results, we have taken the AdaBoost+SVR as the winning Model.

Final model evaluation

After running our model, a minor pre-final step is to be done. We re-invert the scaling in order to obtain values comparable with our target reference. Then both results, and reference values are plotted and compared:

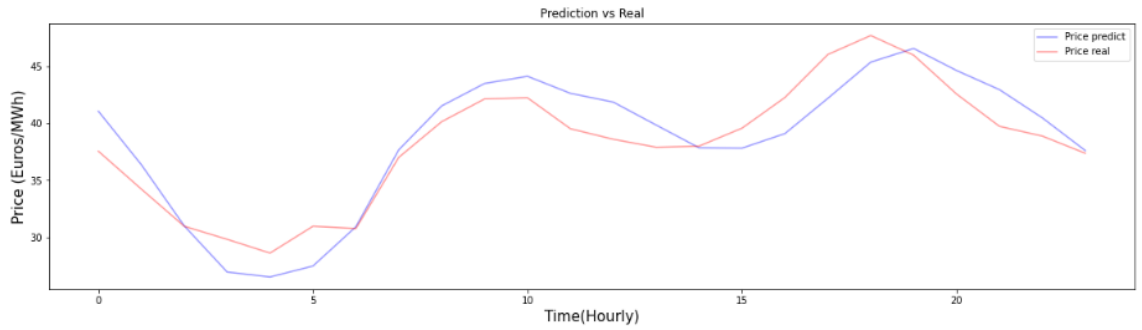

Figure 13. Final model prediction.

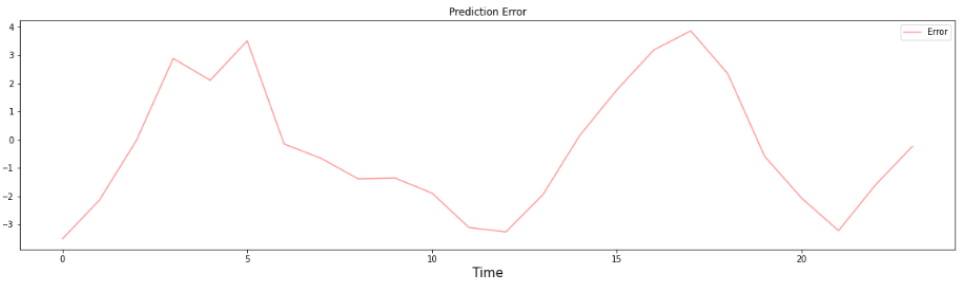

Figure 14. Predicted error.

Conclusion

In this report, six different machine learning models to predict day-ahead electricity prices were validated. Two of the six models we used predicted the 24 hours ahead electricity prices, the Gradient Boosting Regressor was used as the best model at beginning, however according to the metrics we obtained, it showed less accuracy compared to the results from Support Vector Regression model. We also use the SVR model as the basic model for the AdaBoost Regression model to further improving the accuracy of the result.

From the results we found, we noticed that sometimes the complicated model won't always have higher accuracy when it comes to predicting short time period data.

The future work we will do if we have more time on this project is creating a bigger dataset with more attributes and more historical data. Plus, using deep learning methods on energy electricity price prediction is becoming more and more popular, and it shows higher accuracy from the recent paper.

References

Bart De Schutter, Jesus Lago a, b, Ridderb, F. D., Schuttera, B. D., 2022. *Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms*. Applied Energy. Retrieved December 27, 2022, from https://www.sciencedirect.com/science/article/pii/S030626191830196X

Liboreiro J, Filippis A, 2021. *Why Europe's energy prices are soaring and could get much worse.* Euronews.com. *[Article]* Available at: <https://www.euronews.com/my-europe/2021/10/28/why-europe-s-energy-prices-are-soaring-and-could-get-much-worse> *[Accessed]* 24 December 2022

Pierre E, Plantevit M, Robardet C, Tschora L, 2022. *Electricity price forecasting on the day-ahead market using machine learning*. Applied Energy, Volume 313, 2022, 118752, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2022.118752