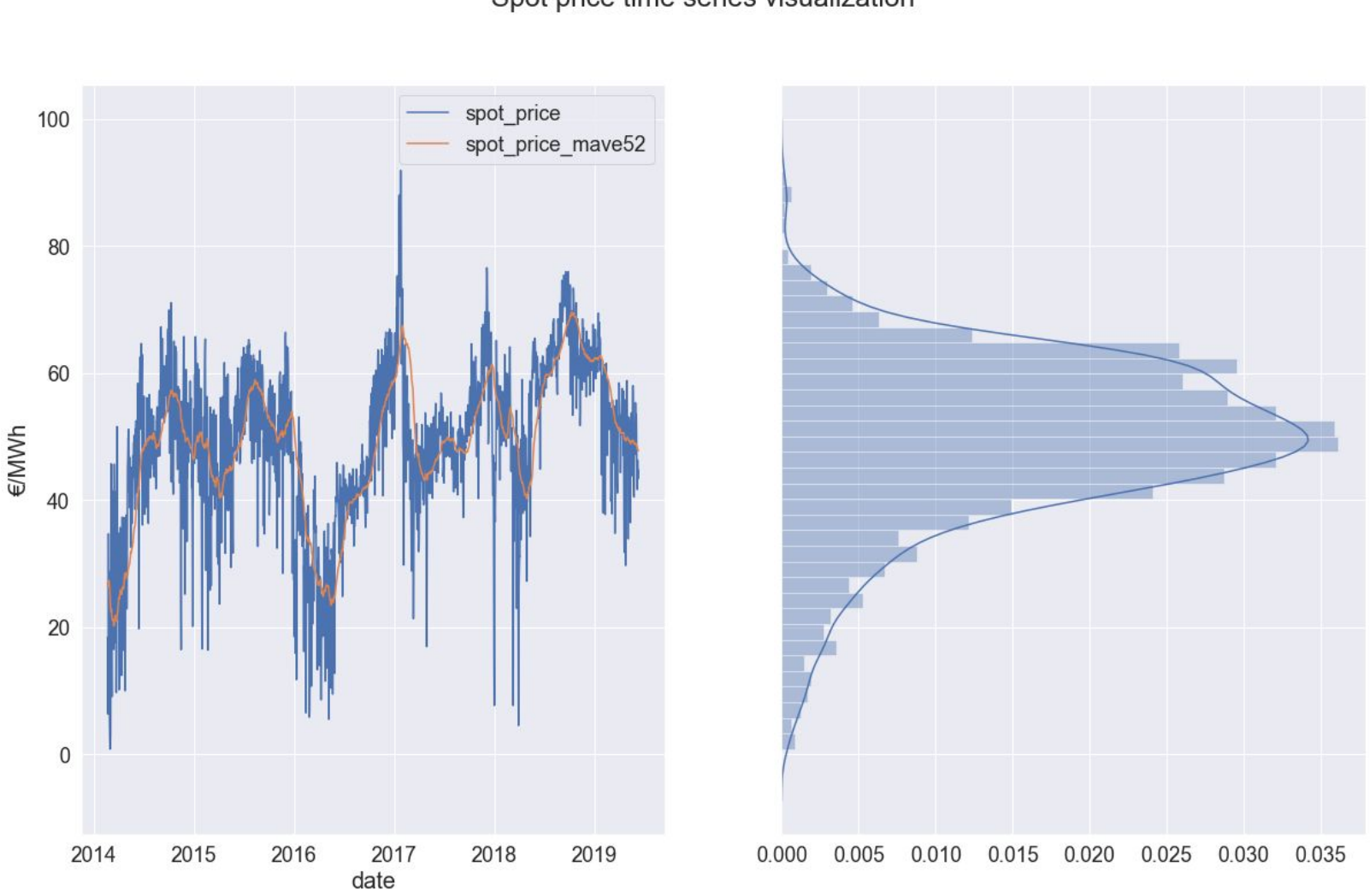
Forecasting Spanish Electricity Spot Price


Spot price time series visualization
spot_price
spot_price_mave52
€/MWh
100
80
60
40
20
0
2014
2015
2016
2017
2018
2019
date
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035


EUA_CO2
coal_API2
NG_TTF
100
80
60
40
20
0
2014
2015
2016
2017
2018
2019
Date

I would like to share how to build a Machine Learning model to forecast spot market electricity prices in Spain.

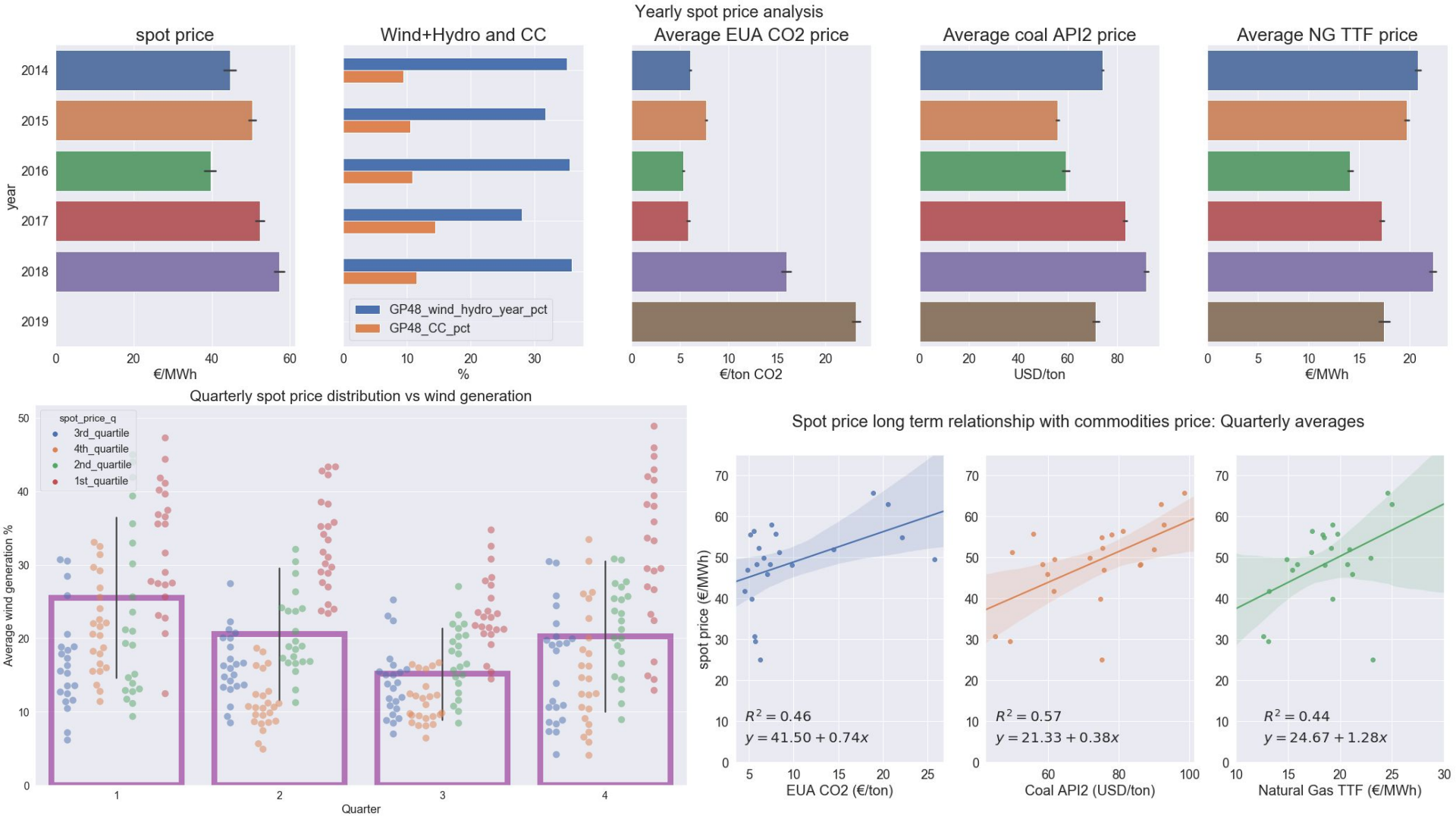
This first step is focused on ingesting data (Python), we will use two sources:

Electricity market data:

- Daily price is set at daily auctions in a pool market. This market is said to be marginal, because there is a matching bid that sets current day price.
- P48 generation by technology: It is also called PHO (Programa Horario Operativo) and it is the definitive schedulled generation program

Financial data:

- EUA (European Emission Allowances, CO2): Important when marginal generator is a coal or natural gas utility
- Coal API2: There are many kinds of coal (antracite, lignite, etc.), so finantial analysis firms deviced this index arranging European data
- Natural Gas TTF: It stands for Title Transfer Facility, it is a Virtual Trading point in the Netherlands, and also a price reference in European HG markets (like Henry Hub or NBP)


Yearly spot price analysis
spot price
Wind+Hydro and CC
Average EUA CO2 price
Average coal API2 price
Average NG TTF price
year
2014
2015
2016
2017
2018
2019
€/MWh
%
€/ton CO2
USD/ton
€/MWh
GP48_wind_hydro_year_pct
GP48_CC_pct
Quarterly spot price distribution vs wind generation
spot_price_q
3rd_quartile
4th_quartile
2nd_quartile
1st_quartile
Average wind generation %
Quarter
Spot price long term relationship with commodities price: Quarterly averages
spot price (€/MWh)
$R^2 = 0.46$
$y = 41.50 + 0.74x$
EUA CO2 (€/ton)
$R^2 = 0.57$
$y = 21.33 + 0.38x$
Coal API2 (USD/ton)
$R^2 = 0.44$
$y = 24.67 + 1.28x$
Natural Gas TTF (€/MWh)

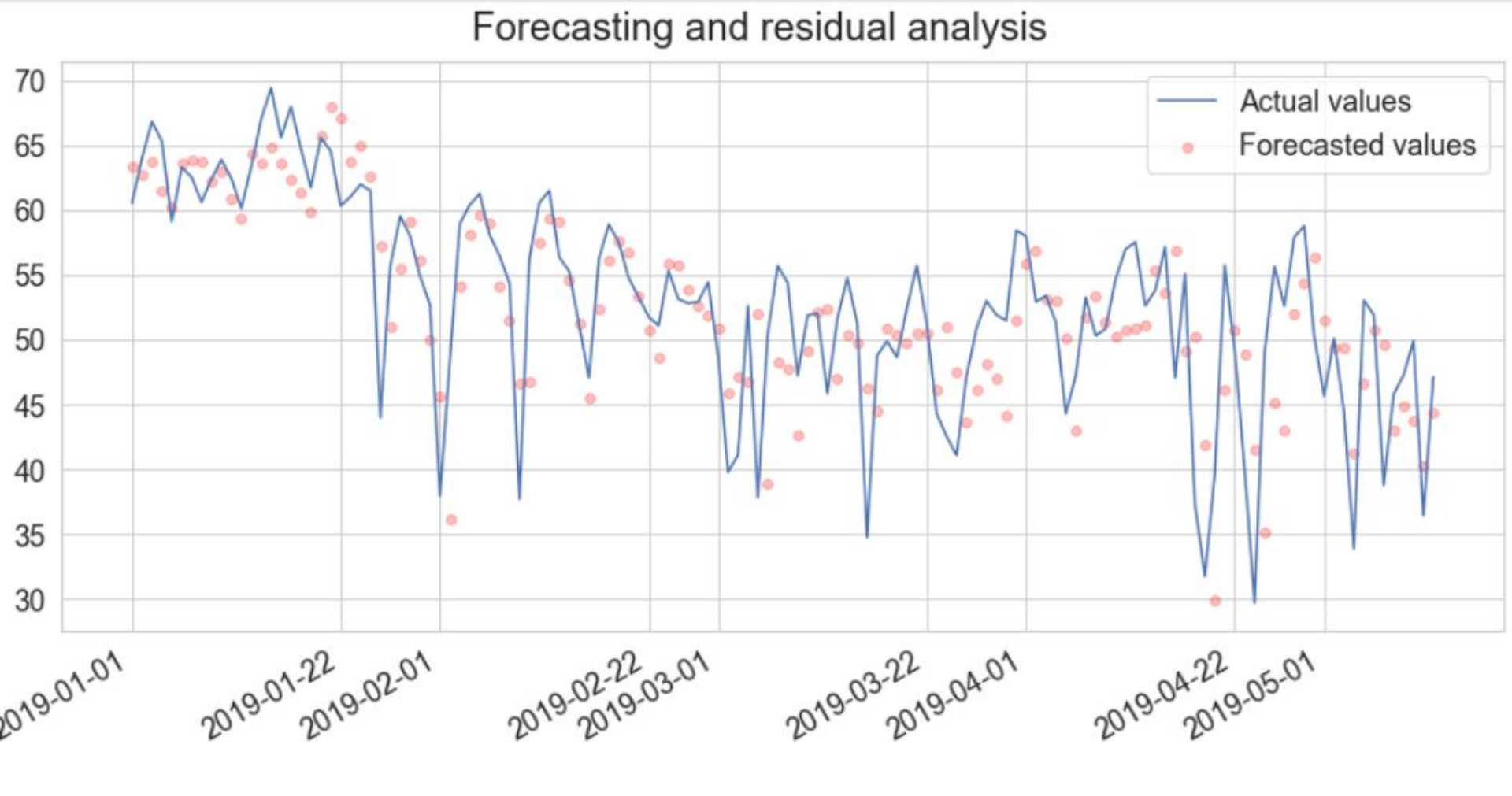
Does renewable electricity production reduce spot market electricity prices in Spain? Why 2018 spot price was so high? Due to fossil fuels or CO2 allowances prices going up?

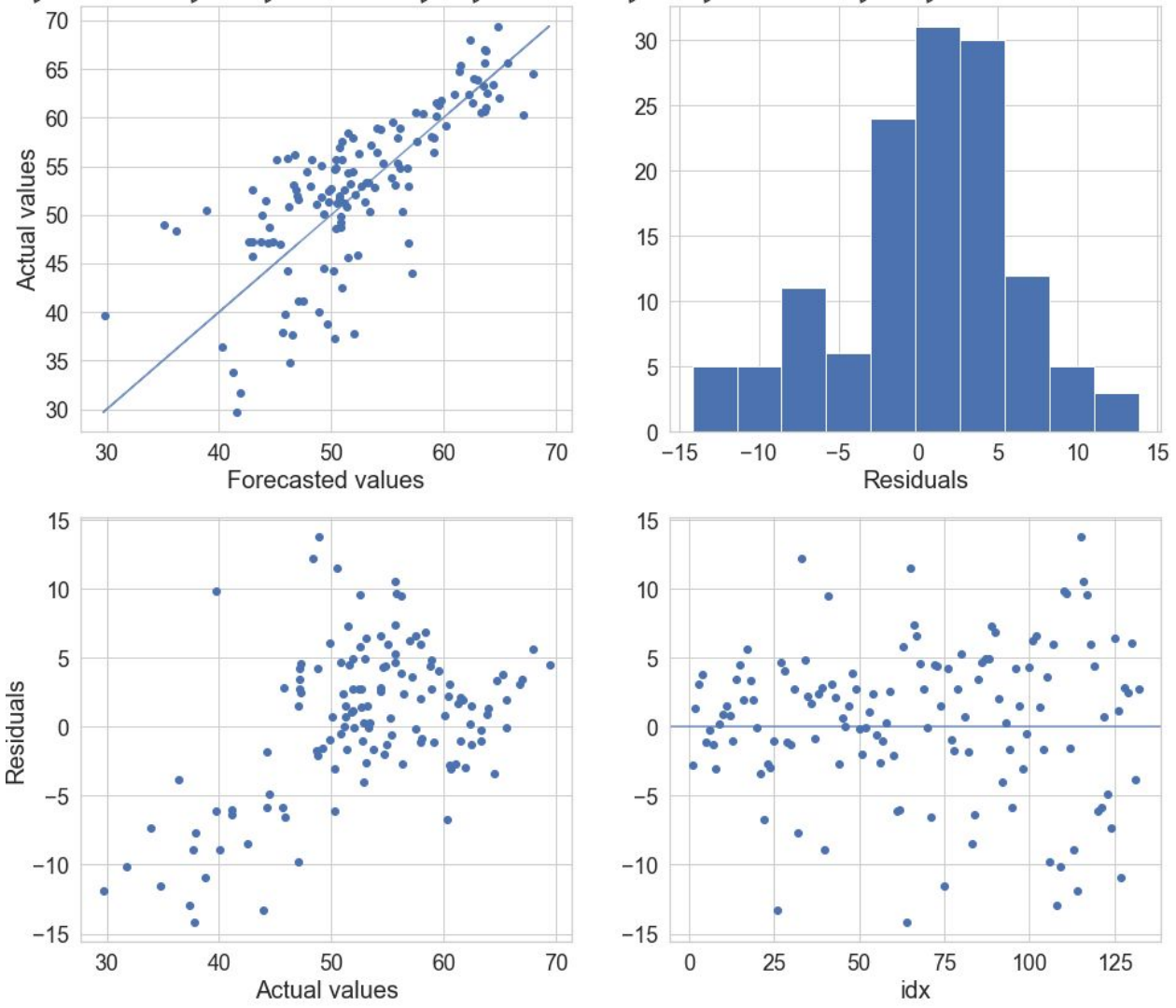
I would like to share an Exploratory Data Analysis (EDA) whose main objective to discover some insights that can be leveraged to build a sound Machine Learning model

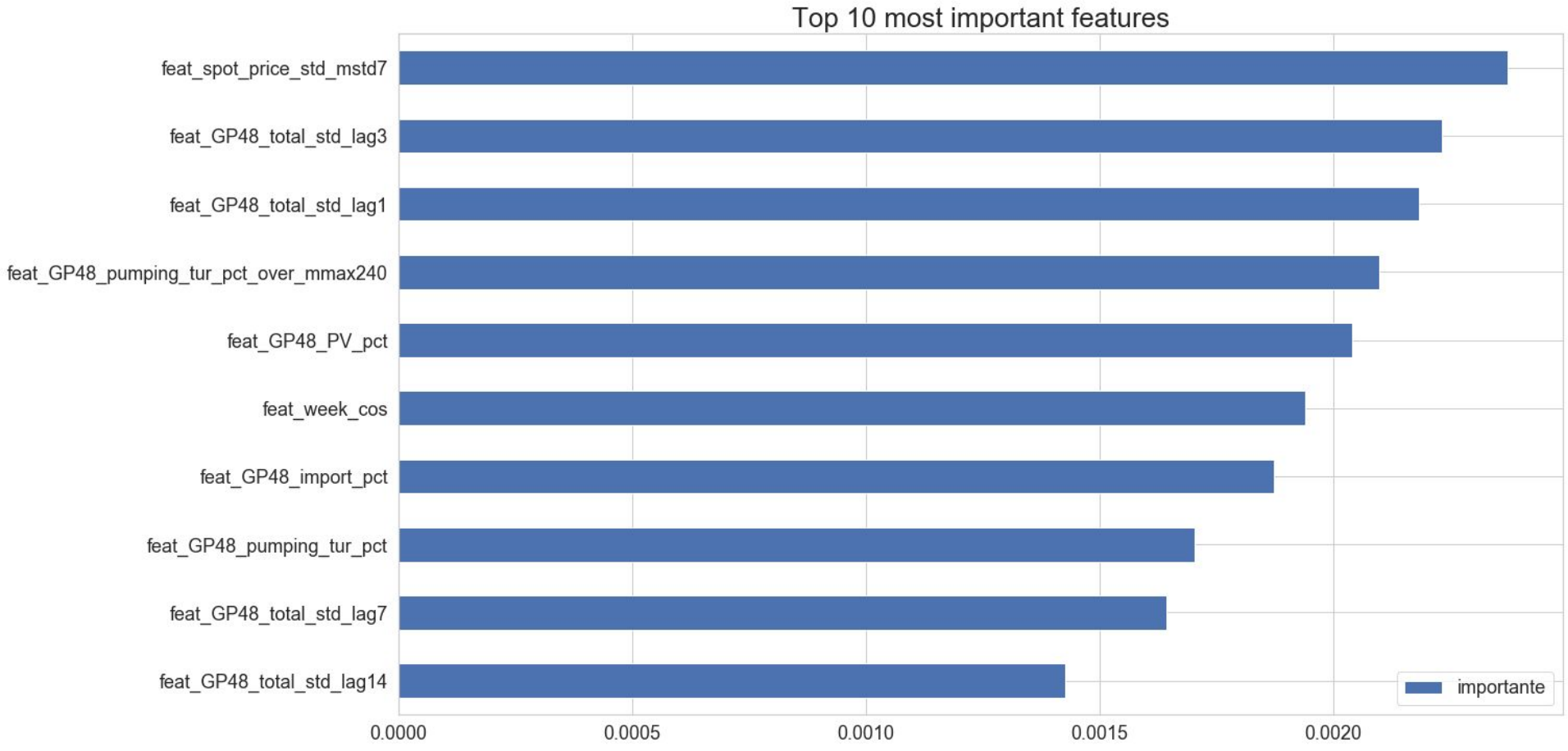
Firstly, I would like to introduce briefly how Spanish electricity market works. Hourly, it is ran an auction, where generators and consumers make their bids. Supply and demand bids are aggregated, therefore it is called a “pool”. It is a marginal auction because bids matching supply and demand are the ones that sets up the resulting price

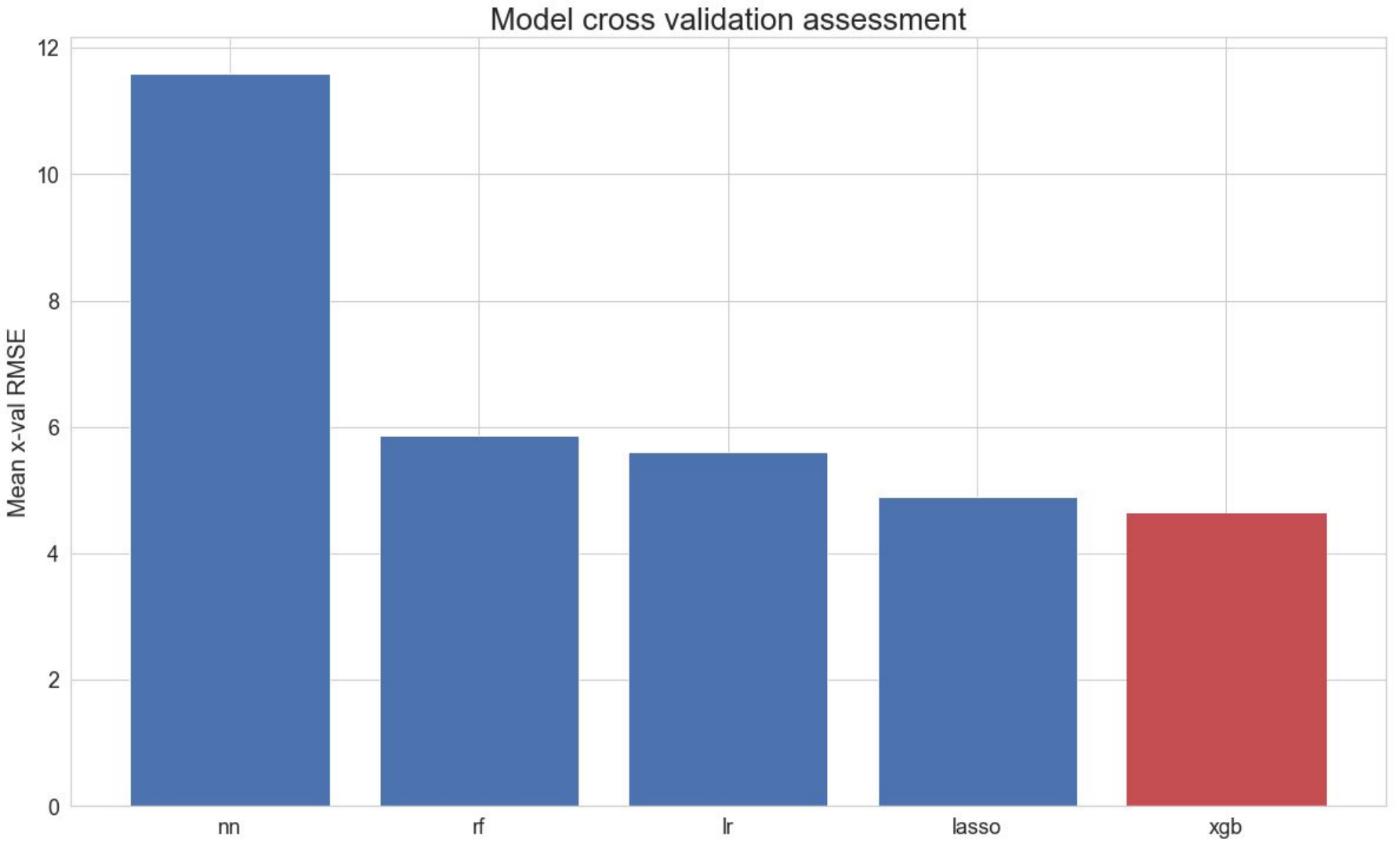
Bidding strategies vary by technology, in general, high CAPEX and low OPEX utilities (like renewables) tend to bid at very low prices; whereas fossil fuels do the opposite. The reason is that, theoretically, they must bid at their variable cost. There are two main exceptions: Nuclear utilities tend to sign bilateral contracts. Hydro utilities bear hidden opportunity costs and may bid higher

These hypothesis are assessed in the EDA notebook. We are ready (and really looking forward) to forecast!


Forecasting and residual analysis
Actual values
Forecasted values
70
65
60
55
50
45
40
35
30
2019-01-01
2019-01-22
2019-02-01
2019-02-22
2019-03-01
2019-03-22
2019-04-01
2019-04-22
2019-05-01


Actual values
Forecasted values
Residuals
Residuals
Actual values
Residuals
idx


Top 10 most important features
feat_spot_price_std_mstd7
feat_GP48_total_std_lag3
feat_GP48_total_std_lag1
feat_GP48_pumping_tur_pct_over_mmax240
feat_GP48_PV_pct
feat_week_cos
feat_GP48_import_pct
feat_GP48_pumping_tur_pct
feat_GP48_total_std_lag7
feat_GP48_total_std_lag14
importante
0.0000
0.0005
0.0010
0.0015
0.0020


Model cross validation assessment
Mean x-val RMSE
12
10
8
6
4
2
0
nn
rf
lr
lasso
xgb

What kind of #machinelearning model will perform better in a forecasting daily electricity spot price?
After computing a comprehensive set of features and assessing how they impact on our objective variable, the next step is to build a set models that can infer this relationship from historical data.

This is a very complex problem: Seasonality, trend, feature interactions and relationships that vary over time. We will start with simple linear regression models and gradually grow to more complex ones: Neural Nets, Random Forest and (of course) eXtreme Gradient Boosting machines) Linear models perform well on this dataset because there are many and good handcrafted features, but finally, XGBs is the best one, with an average percent error metric over 8% on test data (2019). A very good score considering that meteo and water reservoir data is not added and that 2019 is a very particular year. Long live XGBs!