Team 52 - Energy prices forecasting
Project No. 16.
Stakeholder: Conexalab S.A.S

EnergyApp

Bryan Moreno, Andres Bohorquez, Edgard Rodriguez, Andres Ruiz, Carlos Contreras, Victor Loaiza, Antonio Barrios

Business Problem

The main problem to solve in this project is **predicting the cost of the electrical energy offered in the traditional system**. This could help to a reduced rate of return on investment on self-generation and distributed generation projects. The mentioned prediction could be achieved by considering the price that the energy offered in the traditional system can reach during the useful life of the projects and the weather conditions of Colombia.

Business Impact

Having an algorithm that predicts with some precision the cost of the electrical energy offered in the traditional system would make more efficient the investment in self-generation and distributed generation projects. This will have an environmental impact since it makes self-generation and distributed generation projects more accessible and attractive. These projects contribute to the reduction of greenhouse gases, eliminate the environmental impact of energy transport and minimize the need to intervene in water sources for the energy generation. In addition, it will allow the government to take decisions to promote and encourage the use of sources of unconventional energy.

Data

The information provided by Conexalab SAS is:

- The price of the kWh of energy on a monthly basis for 5 energy marketers, for a period of 8 years (2013 - 2020)

Nevertheless, this data was not sufficient for the purposes of the project. Therefore, we look into external sources and variables such as:

- Weather conditions: precipitations, temperature, sun irradiance. (Ideam)
- Supply and demand of energy for a given period of time.
- Hydrological factors.
- Energy exports and imports.
- Energy production mix (by production method).

Methods

Visualization

To understand the business context behind the energy price prediction, we propose the historical visualization of prices, through univariate and bivariate analysis, the prediction proper. We can provide some visualizations such as:

- Correlation plots with other variables which we can find from other sources.
- Autocorrelation and partial autocorrelation plots to search for seasonal patterns.
- Prices of energy per marketer over time, including a near month forecast and its confidence interval.
- Other variables (weather conditions, fuel prices,..) magnitudes over time
- Residual plots for model accuracy evaluation

Models

Supervised Learning

We start by doing an exploratory data analysis on the databases obtained from XM to determine which variables have the most impact on the cost of the electrical energy . Then we will do some preliminary tests with different regression methods, for example linear regression, arima, sarimax and we will implement a Machine Learning regression. From those we will choose the better in both accuracy and scalability.

Time series analysis

Starting from the data provided by Conexalab, time series analysis can be executed to provide a forecast within monthly periods. Exploratory data analysis may include trend analysis, search for seasonal patterns and their corresponding hypothesis testing. We can apply forecasting models such as AR, ARIMA, exponential smoothing and dynamic regression, whose implementation is contained in the statsmodels python library and are documented in books like [1].

Depending on the additional data we are able to collect, including weather conditions, oil and coal prices, supply and demand of energy, we can use more complex models such as ARIMAX and Neural Networks based models. Those models are well documented for the energy forecasting problem, as in the review article [2], and have shown to be well suited for the problem we deal with.

Interface

The visualization will be achieved by creating a control panel with python technologies, specifically with libraries like Dash. On the dashboard, the user will be able to observe historical price analysis and will be able to interact to obtain the predicted future price of energy.

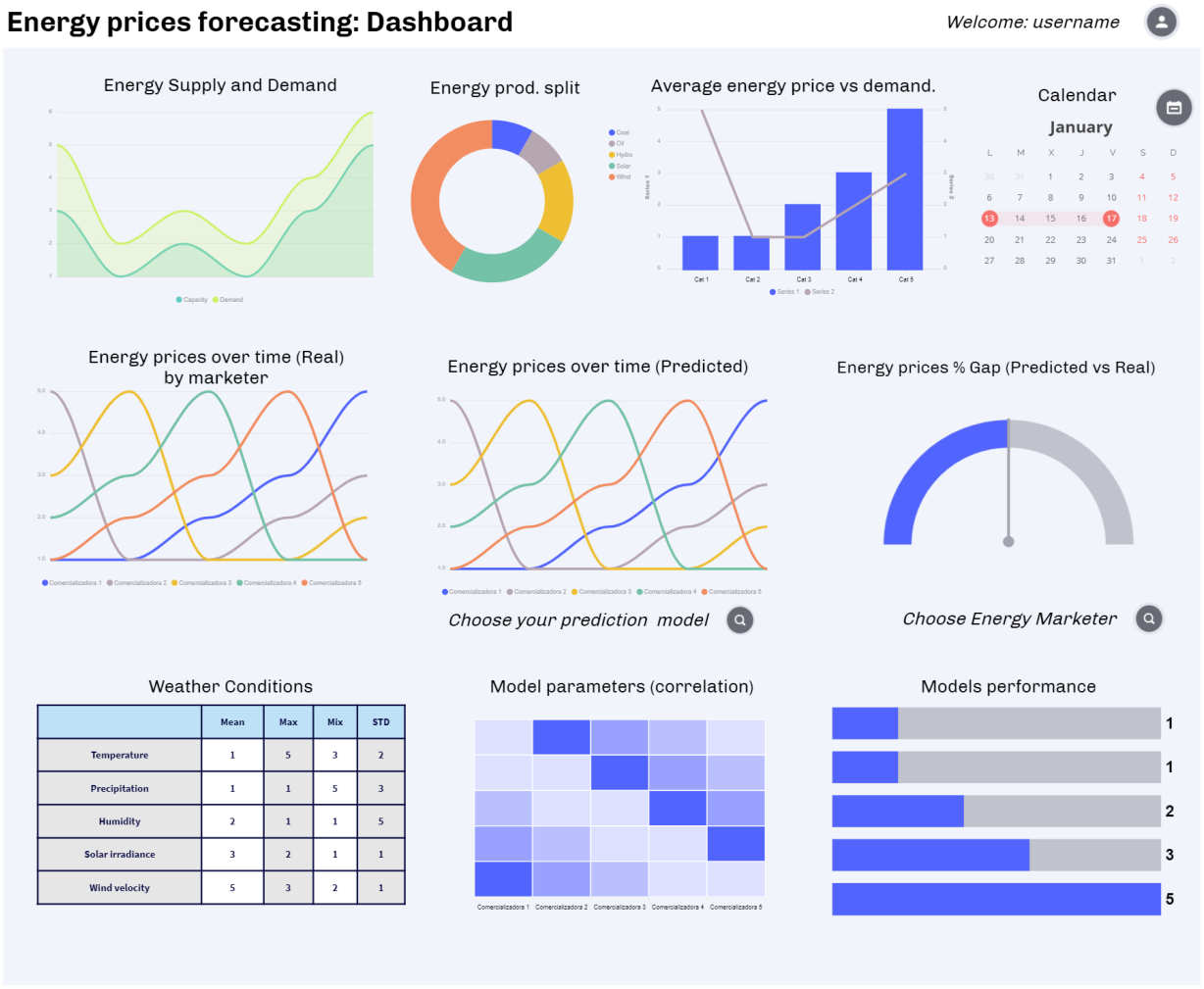


Figure 1. energy prices forecasting:dashboard

Milestones

Here we provide details on the milestones we intend to achieve in our project. In particular, we have outlined four different versions: we expect to finish Version 1.0 with 100% probability, Version 1.3 with ~70% probability, Version 1.6 with ~20% probability, and Version 2.0 with ~5% probability (if things go extremely well)

Version 1.0: Build a simple dashboard with a few plots of the data that we found to be relevant in the cost of the electrical energy and a time series predicting the future costs.

Version 1.3: Build a prediction model based on the selected model to get the predicted values of energy.

Version 1.6: Add an interactive plot based on the results of the prediction model.

Version 2.0: Add the option to manually input new data to update the databases. Implementing state of art advanced models.

Date	Deliverable	Details
Week 1	Data Sets available	Complete definition of the datasets that would be used for the project, datasets ready to go
Week 2	Dataset Cleaning	Clean an organized datasets, including outlier drop, imputation and agroupation
Week 3	Exploratory Data Analysis/insights into ML and other prediction tools	Exploratory Data analysis using jupyter to gain the core insights of the data and the behaviour of it
Week 4	Project infrastructure designed and running/ FrontEnd design advance	Defined and running project infrastructure, be it aws, or something similar,including web server, data holder and security groups. Basic frontend
Week 5	Upload of project to cloud	EDA and other algorithms working on cloud
Week 6	Project migration to cloud	Project already running on cloud with front end working and dashboards ready
Week 7	Final touches	Last minute bugs and problems solved, first insights on project presentation
Week 8	Project Presentation	finished and practiced project presentation

Concerns

Our biggest concern at the moment is the database provided by conexalab, since the database only consists of a excel file of 4kbs containing 60 rows. With this really small amount of data it is not possible to implement complicated models like machine learning, SVR, XGBoost etc...

Even though we can get plenty of data from XM public website, in the best case scenario we will be able to do a time series or a simple linear regression. To face this problem we hope to meet Conexalab as soon as possible to see what other information we can get from them and what other tools they can offer us and then decide if this project is actually feasible or not.

Also, our team lacks enough experience with the Colombian energy market. Therefore, the search for information to select the adequate variables task would require much time and we might not find enough insights in the available time.

Other smaller problems are that in our team we have nobody with knowledge on the front-end so to cope with this we will follow correlation one advice and stick with dash to do it.

Data Description

The information provided by Conexalab SAS is:

- The price of the kWh of energy on a monthly basis for 5 energy marketers, for a period of 8 years (2013 - 2020)

Nevertheless, this data may not be sufficient for the purposes of the project. Therefore, we looked into external sources such as:

Dataset Name	Description	Provider	Source
Energy Demand	Historical demand per day in kWh in Colombia. Start date: 2000 - 01 - 01 End date: 2020 - 12 - 31	XM	http://portalbissrs.xm.com.co/dmnd/Paginas/Historicos/Historicos.aspx
Energy Supply	Historical generation per day in kWh in Colombia. Start date: 2000 - 01 - 01 End date: 2020 - 12 - 31	XM	http://portalbissrs.xm.com.co/dmnd/Paginas/Historicos/Historicos.aspx
Energy price	Historical stock price per	XM	http://portalbis

on the stock market	Generator and day in COP in Colombia. Start date: 2000 - 01 -01 End date: 2020 -12 - 31		srs.xm.com.co/trpr/Paginas/Historicos/Historicos.aspx?RootFolder=%2Ftrpr%2FHistricos%2FBolsa&FolderCTID=0x012000394993FA303733428C33EC91D1DFA6DB&View=%7B5CA2173E%2D1541%2D4EC7%2D9D1C%2DE145E3DFFAE3%7D
Hidrology	Reservoirs and other hydrological variables in Mm3 and kWh. Start date: 2000 - 01 -01 End date: 2020 -12 - 31	XM	http://portalbissrs.xm.com.co/hdrlg/Paginas/Historicos/Historicos.aspx
Multivariate ENSO Index	El Niño–Southern Oscillation (ENSO) is an irregular periodic variation in winds and sea surface temperatures over the tropical eastern Pacific Ocean, affecting the climate of much of the tropics and subtropics. It represents the level of precipitation around the pacific countries. Start date: 1979 - 01 End date: 2021 - 06	(NOAA) National Oceanic and Atmospheric Administration	https://psl.noaa.gov/enso/mei/data/meiv2.data
Colombian coal prices	Yearly energy prices of colombian coal used for energy generation	Index-Mundi	https://www.indexmundi.com/es/precios-de-mercado/?mercancia=carbon-colombiano&meses=120&moneda=cop#google_vignette

Exploratory Data Analysis

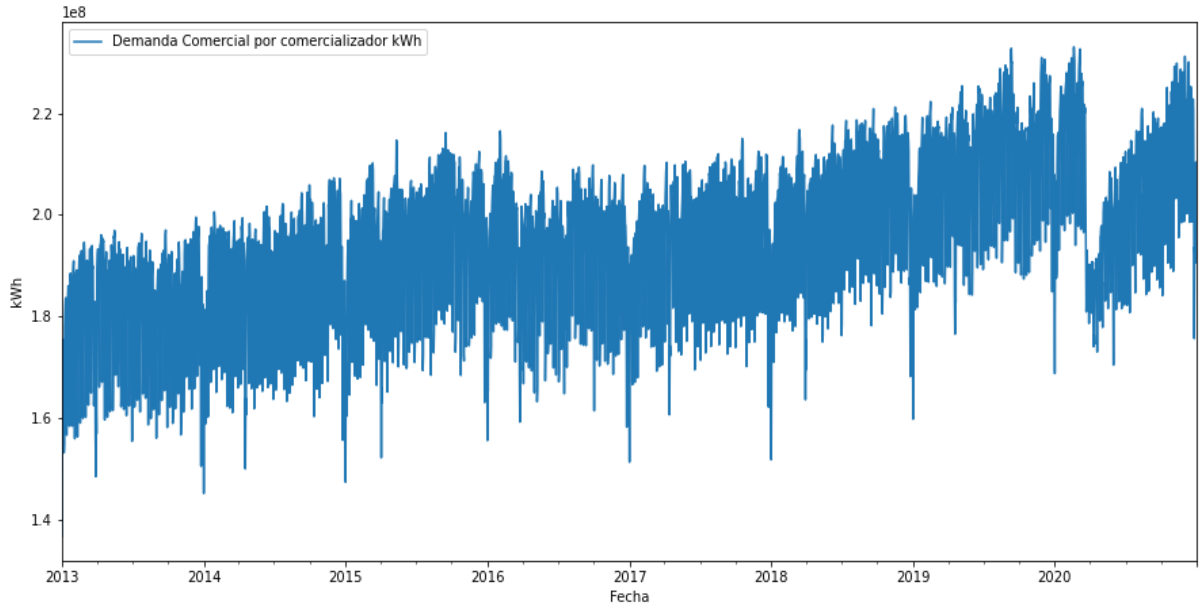


Figure 2. *Energy demand vs time: Daily from 2013 to 2020*.

From figure 2 we observe an increasing trend regarding the energy demand through time, it seems that the annual mean increases for every year, especially for 2018 and 2019.

For each year there is also a short period where the demand decreases, coinciding with the beginning of the aforementioned year. This can be due to the fact that these are the most common dates for holidays in Colombia and many industries may stop their production.

Additionally, during 2020 there is a steep decrease in the energy demand, which coincides with the application of measures related to the COVID-19 pandemic.

To confirm the visible trend, we ran a non-parametric Mann-Kendall test, which with a p-value < 0.05 corroborates our belief of an **increasing trend.**

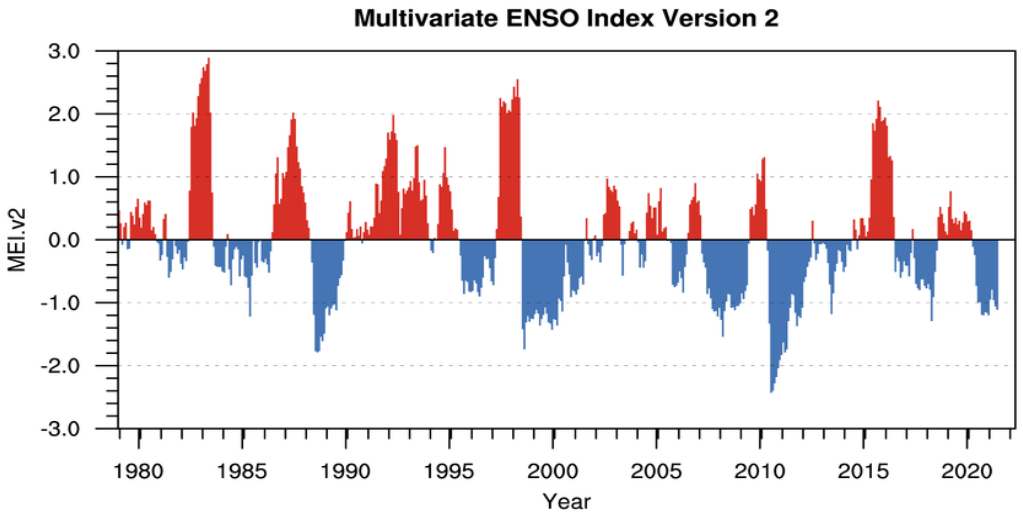


Figure 3.*Multivariate ENSO Index - MEI: Monthly from 1979 to 2021*.

The El Niño/Southern Oscillation (ENSO) - a naturally occurring anomalous state of tropical Pacific coupled ocean-atmosphere conditions - is the primary predictor for global climate disruptions. These can persist over several seasons and thereby produce severe regional effects, especially in the countries located around the Pacific Ocean. Since most of the energy in Colombia is generated with hydroelectric, annual rainfall is closely related to energy generation capacity. The MEI, which combines both oceanic and atmospheric variables, facilitates in a single index an assessment of ENSO. It especially gives real-time indications of ENSO intensity, and through historical analysis - provides a context for meaningful comparative study of evolving conditions.

The Multivariate ENSO Index (MEI.v2) is the time series of the leading combined Empirical Orthogonal Function (EOF) of five different variables over the tropical Pacific basin:
sea level pressure (SLP)

- sea surface temperature (SST)
- zonal components of the surface wind
- meridional components of the surface wind
- outgoing longwave radiation (OLR)

The index value indicates the occurrence of the single climate phenomenon fluctuation between three phases:

- MEI > 0.5: El Niño
- 0.5 <= MEI <= 0.5: Neutral
- MEI < -0.5: La Niña

Key features in Colombia of composite positive MEI events (warm, El Niño) include decrease in rainfall in relation to the monthly historical average and increase in air temperatures, especially in the Caribbean and Andean regions. Key features of composite negative MEI events (cold, La Niña) are of mostly opposite phases.
Taking this into account, the figure above shows some extreme events that Colombia has had on the recent years:

- El Niño: 1982, 1997, 2015
- La Niña: 1988, 1998, 2010

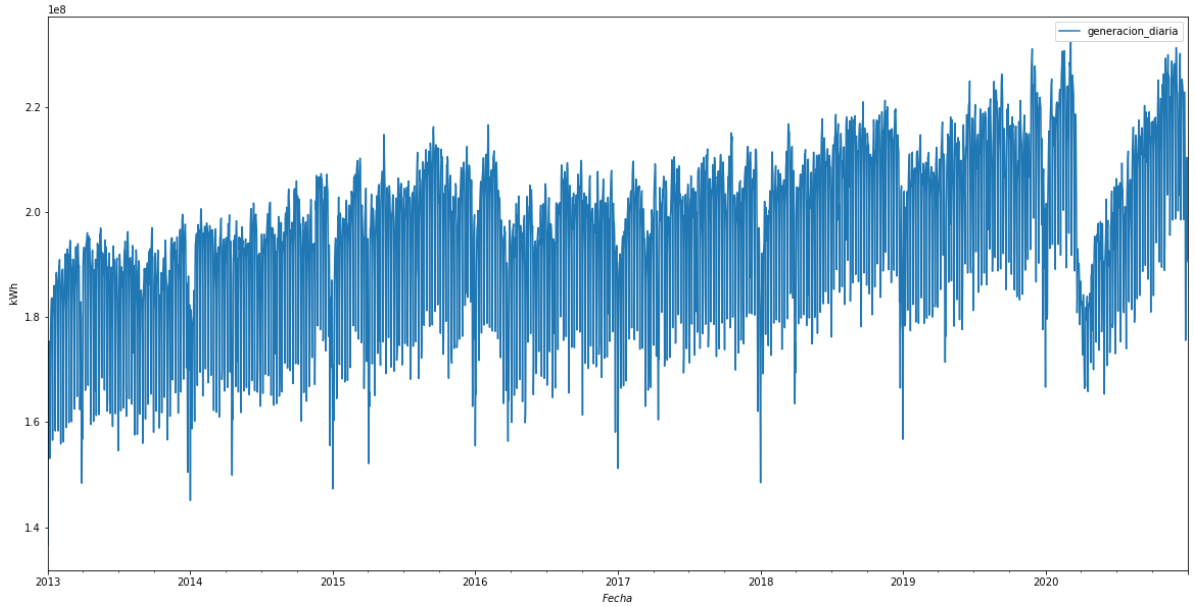


Figure 4. *Energy Generation vs time: Daily from 2013 to 2020*.

We observe almost the same behavior in figure 4 and figure 2, which is expected since in a vital service like the energy the offer would try to fulfill the demand as much as possible.

Here in figure 5 we see that the Hydroelectric sector takes around two thirds of the total market, while the termoelectric sector is in second place with near the last third and the other sectors make an almost negligible contribution to the national generation. This information agrees with the information we got from our talks with the representatives of conexalab.

The termoelectric sector generation grows whenever the hydroelectric falls down as expected from the two main players of the national market. On 2016, it seems to be a weird behavior with a suddenly prolonged drop of the hydroelectric sector, this drop correlates with the 2016 peak of figure 3 which indicates that the El Niño phenomenon that happened on this year with an intensity so strong that has not been seen on at least 20 years had a heavy effect on the market.

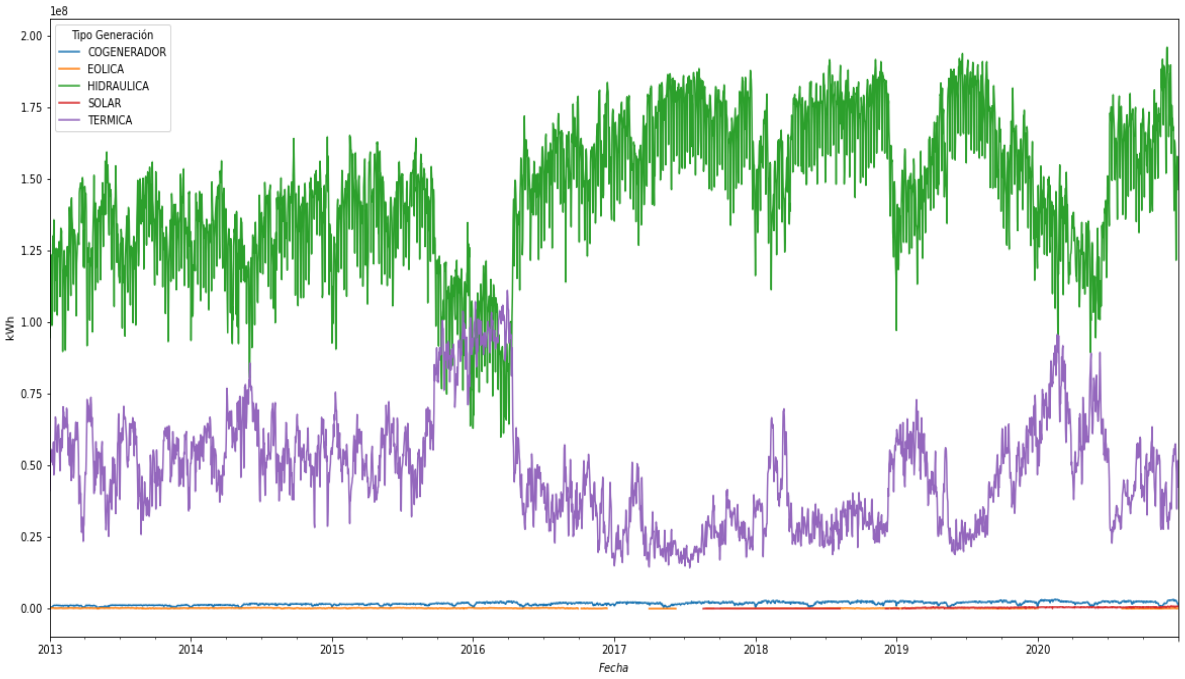


Figure 5. *Energy Generation per generator type vs time: Daily from 2013 to 2020*.

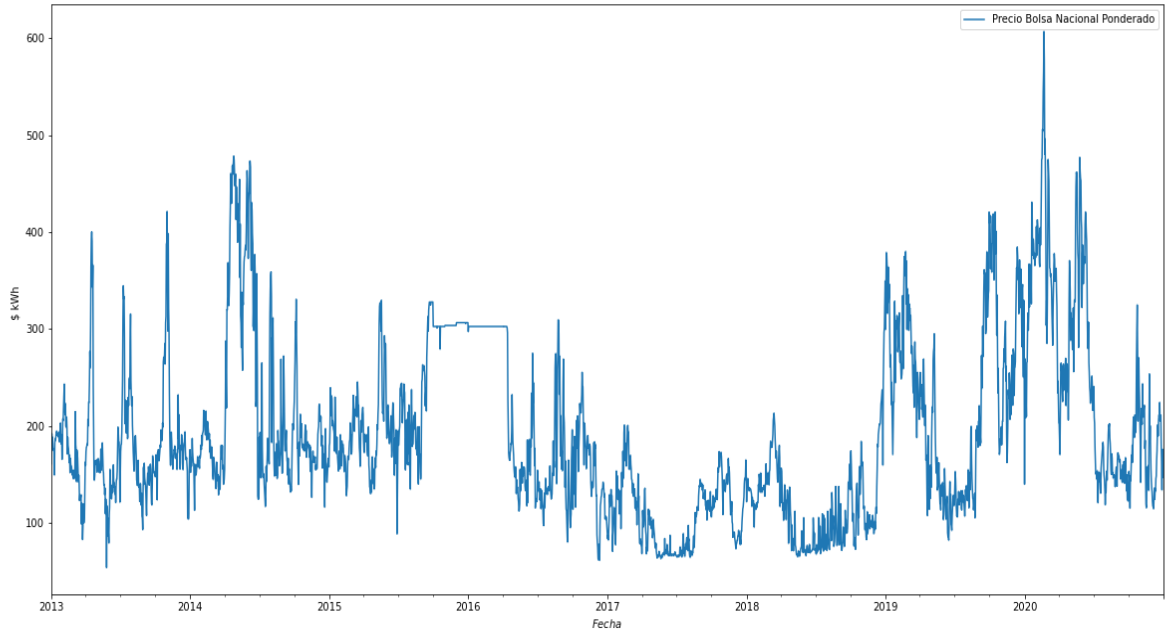


Figure 6. *Energy price per kWh vs time: Daily from 2013 to 2020*.

On figure 6 we see the daily price of energy per kWh, we see in the 2015-2016 range that the prices are almost constant, this region correspond to the ‘fenomeno del niño’ found in figures 5 and 3, the almost constant value is probably due to government intervention during those times. Another interesting fact to note is that there are peaks in the price whenever the hydroelectric generation drops so we expect a highly negative correlation between these variables.

We also see some oscillations and a peak around march of 2020 which correlates with the drop in the energy demand in the same date. This seems to relate with the application of measures related to the COVID-19 pandemic.

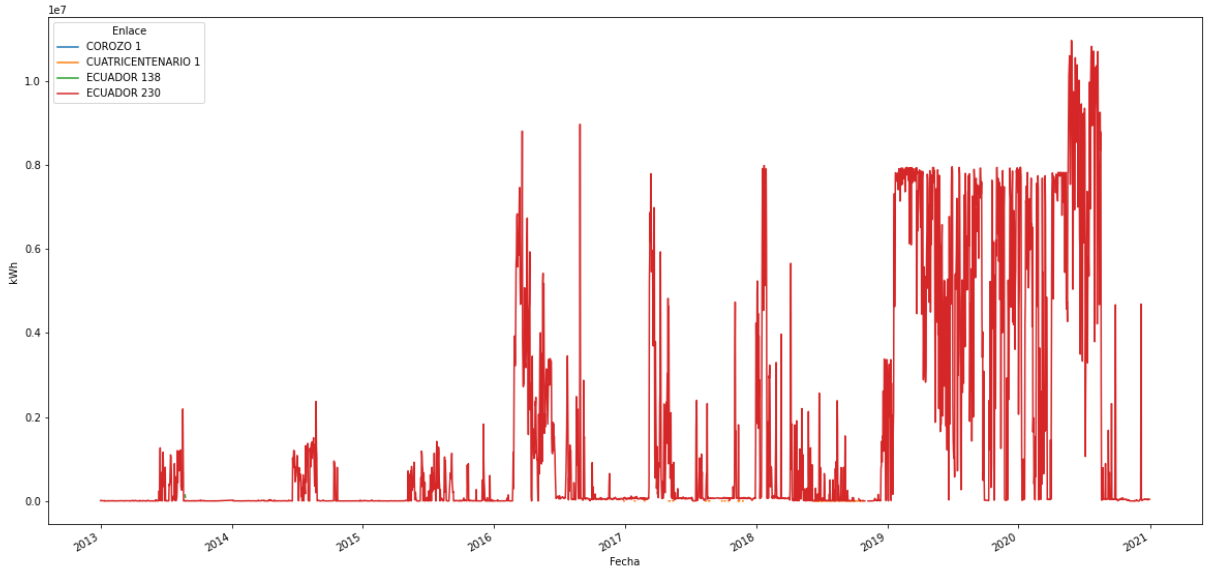


Figure 7. *Energy importations kWh vs time: Daily from 2013 to 2020*.

We see small regions of peaks of importation around 2016, 2017, 2018 and a large region of peaks from 2019 to almost 2021 which are times where the hydroelectric sector drops production, so it seems that Colombia uses energy from Ecuador to supply its demand.

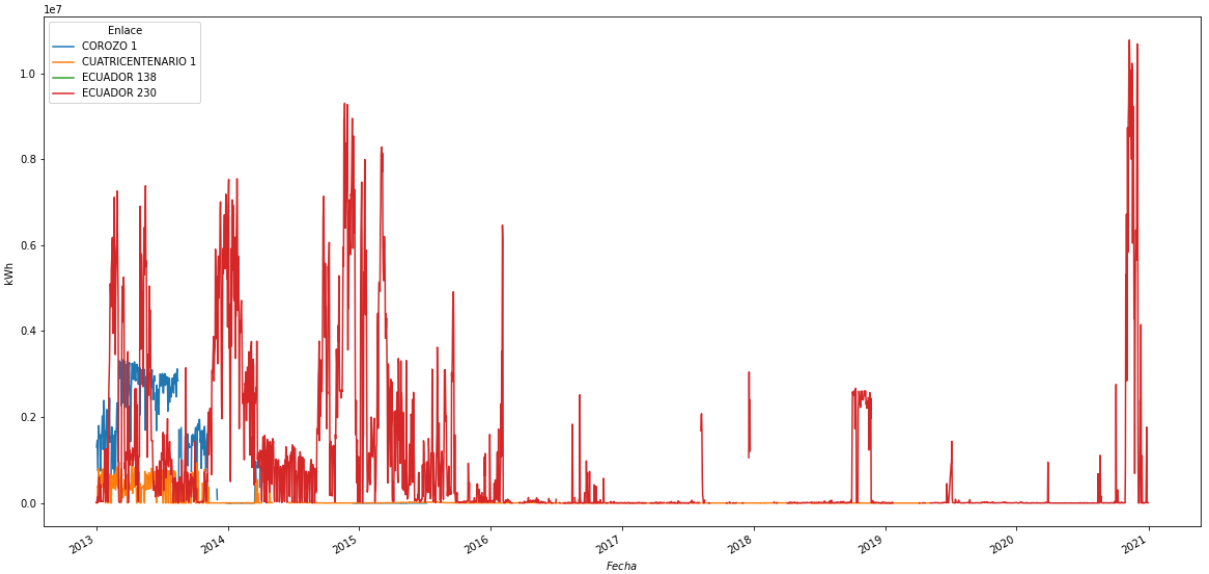


Figure 8. *Energy exportations kWh vs time: Daily from 2013 to 2020*.

From figure 8 we see that colombia used to export energy quite frequently, but stopped on 2016 on the ‘fenomeno del niño’, it seems that it started to export again on the end of 2020,

it is recommended to see data from 2021 to see if this increase was temporary or if it is a new trend.

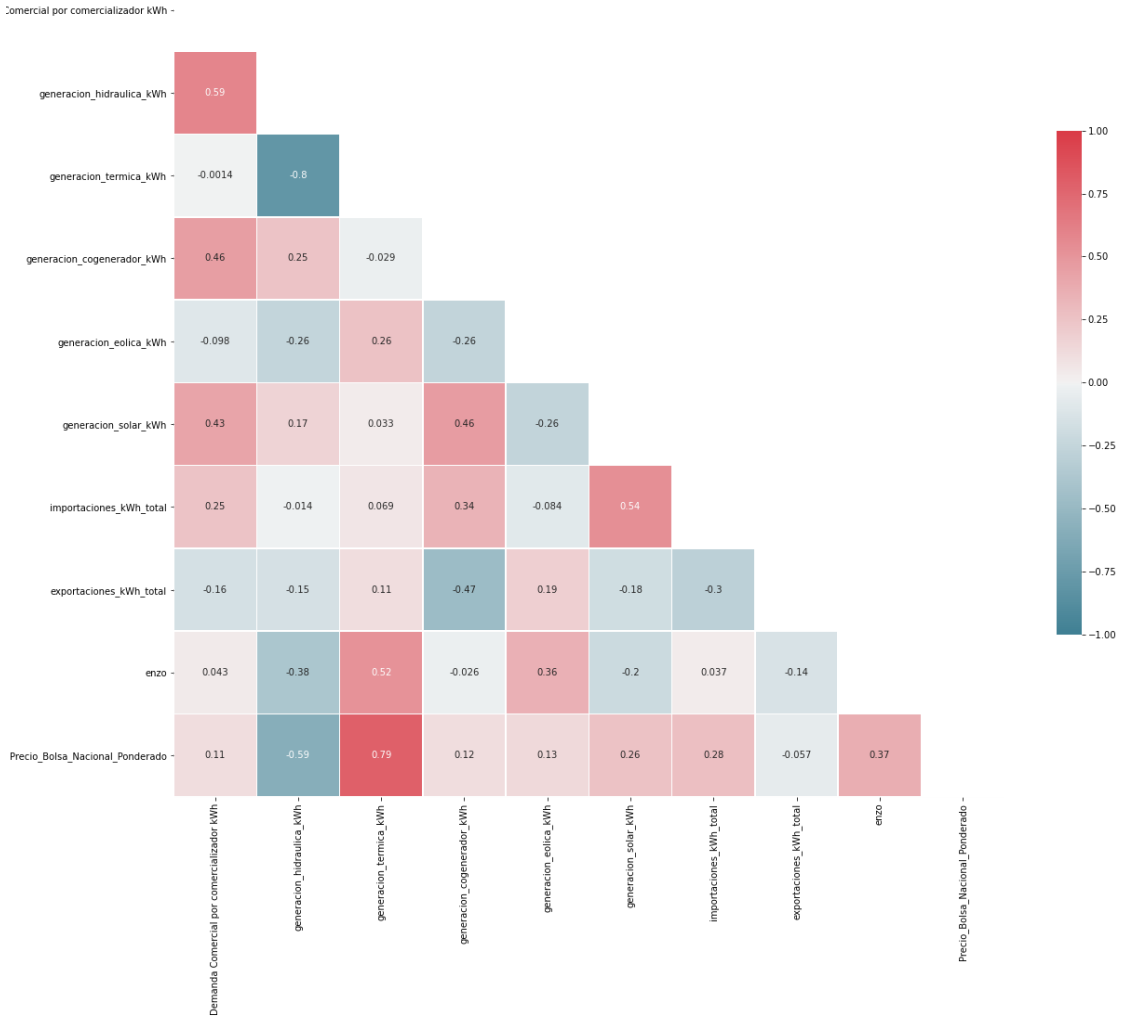


Figure 9. *Variable correlation matrix*.

Looking at the normalized correlation matrix the most correlated variables to our target variable are (Precio_Bolsa_Nacional_Ponderado):

- generacion_termica_kWh
- generacion_hidraulica_kWh
- enso

This makes sense according to the previous knowledge we had: the price is subject to offer and demand, so the quantity of energy produced (our two first variables) will be a driving force into its definition.

Also, “El niño” phenomenon, depicted by the “enzo” variable, is also correlated with this price. The correlation coefficient doesn't seem to be so strong but it can be due to factors

such as the granularity of our data: the data we have for this climate phenomenon is monthly and all of our other variables are daily, so to calculate this coefficient, the “enzo” variable is interpolated throughout the month with an step interpolation.

To **conclude** this preliminary EDA, we had as goal choosing the most relevant variables to predict **price.** To do this we ran a Granger causality test, which will tell us whether or not a time series can be useful to predict another time series.

The steps we took to correctly use this test were:

1. **Normalizing:** We normalized our variables using the mean and standard deviation.
2. **Detrending:** We detrended our variables to avoid having issues when calculating the correlations. We detrended by differencing, constructing a new series where the value of the new series at a time “t” is the difference between the observation at a time “t” and a time “t-1”

Not surprisingly, the time series which proved useful (p-value<0.05) to predict our price time series were the 3 most correlated that we identified via our correlation matrix:

- **generacion_termica_kWh**
- **generacion_hidraulica_kWh**
- **enzo**

Models

To evaluate our models we use the *symmetric mean absolute percentage error* sMAPE metric given by:

$$sMAPE(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\frac{|x_k - y_k|}{(|x_k|+|y_k|)/2}$$

The main reason to use this metric over other metrics is its easy interpretability, being dimensionless allows it to compare different datasets and is symmetry which plays a crucial part of the machine learning algorithms by reducing bias.

Also sMAPE is selected instead of the more traditional *mean absolute percentage error* (MAPE) metric because of the issues that affect MAPE [3,4].

ARIMA

We use an ARIMA model, short for ‘Auto Regressive Integrated Moving Average’ which is a popular model that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

For the train set we used 80% of the energy price data (2000-2016) and the last 4 years of data was used as a validation test. In figure 10 we see the plot for the validation test, here our predictions overlaps with the data quite nicely and we got a value of 8.6 for sMAPE which is quite good.

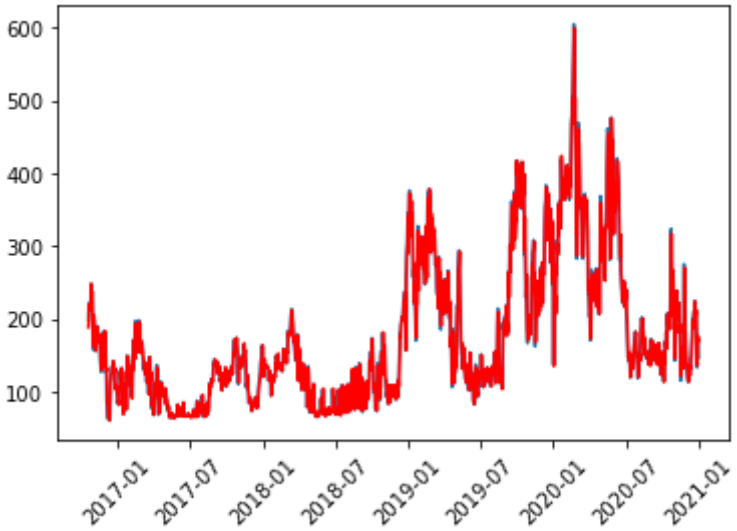


Figure 10. *Arima prediction (red) and* validation test (blue) for daily predictions.

However these were day by day predictions, when we try to make long term predictions we get really awful results, the prediction gets stuck on a constant value and does not change after a few days.

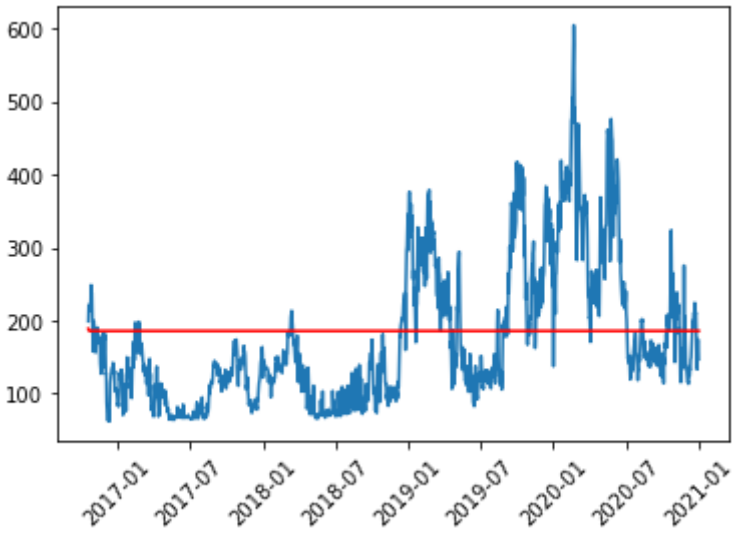


Figure 11. *Arima prediction (Red) and* validation test (Blue) for long term predictions.

We got this kind of behavior with other models (DNNs, LTSM and SARIMA) where we predicted quite good the prices in the validation set day by day but on long term predictions the models tend to fail and only give constant values so we omit them in this writing.

LSTM-DNN model

This is a hybrid forecaster combining an LSTM and a DNN network. The motivation behind this hybrid structure is to include a recurrent layer that can learn and model the sequential relations in the time series data as well as a regular layer that can learn relations that depend on non-sequential data.

This model was already used in [5] for daily and hourly predictions, there the authors compared over 27 other common approaches for day-ahead electricity price forecasting and was shown to obtain a predictive accuracy that is statistically significantly better than all other non ML models and was the second best among the ML models they tried.

We choose this model over the other ML models on the paper since this has way better mid and long term forecasting than the others and shows decent long term predictions if more variables besides the energy demand are added (for example enzo, hydroelectric generation, thermal energy generation, Offer price etc.)

The model uses a DNN to process the regular inputs and an LSTM to process the time sequences . Then, the outputs of these two networks are concatenated into one vector and this vector is fed into a regular output layer.

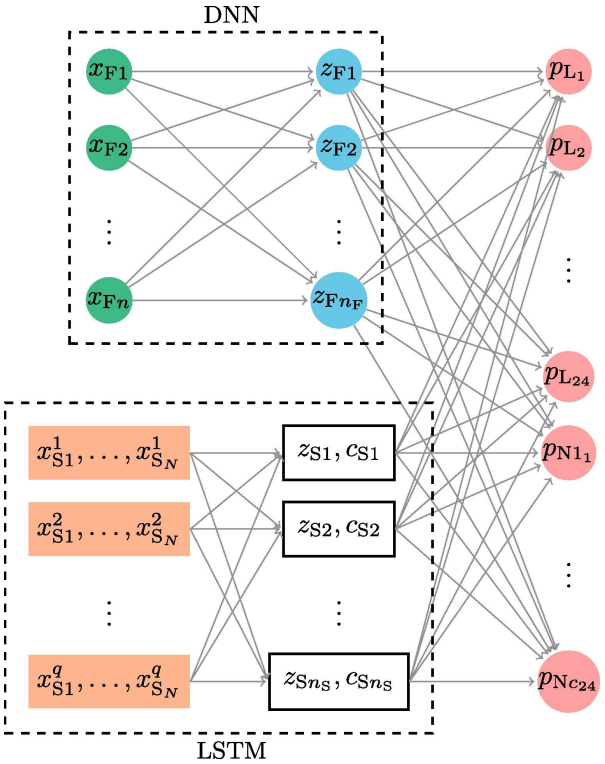


Figure 12. *Hybrid DNN-LSTM network* [5].

We train the model with 184 neurons for the LSMT part and 40 for the DNN part, training during 22 epochs with early stopping. This model takes at input the previous 14 days values of the energy price and feed it on the LSTM and the previous day energy demand for the DNN

Day ahead prediction

Here we show the day ahead prediction,we got sMAPE values of 16.5 and 23 for the training and test sets, these values can be greatly improved by training more the model and having more neurons in the DNN part (we were able to get up to 8 sMAPE on the training and test sets). However, doing those changes greatly damages the long term predictions.

The sMAPE values of 16.5 and 23 may seem high at first glance however they are in the range 12-23 that [5] found in their 27 models predicting the energy price of Belgium and France so they are still in the range of precision of the frequently used models in this kind of problem.

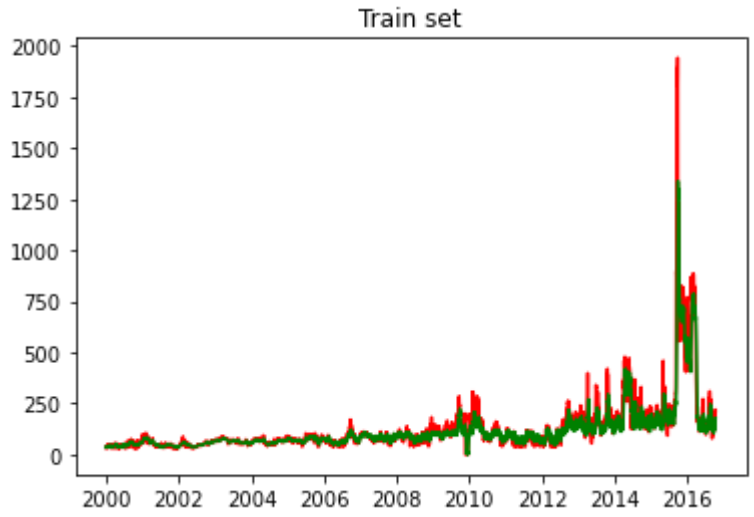


Figure 13. *Hybrid DNN-LSTM network prediction (green) and* training set (red) for daily predictions.

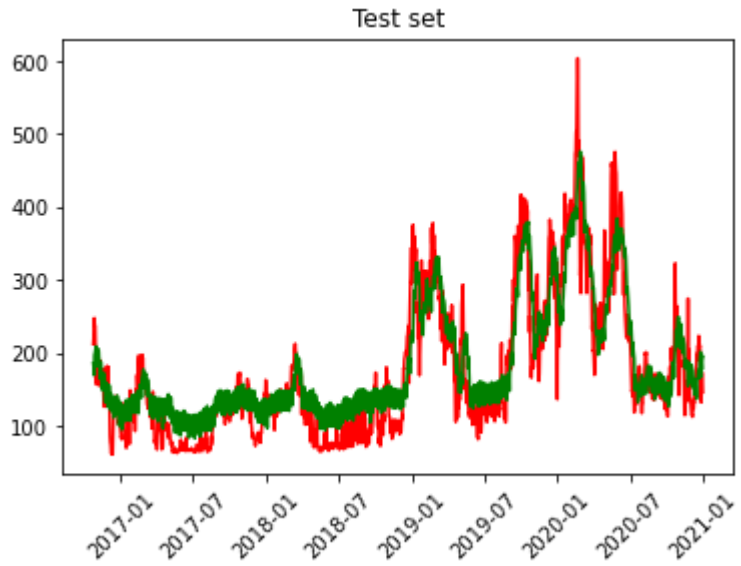


Figure 14. *Hybrid DNN-LSTM network prediction (green) and* test set (red) for daily predictions.

Long term predictions

For the long term we use the output prices of our model as input and take the current value of the energy demand for the next predictions, we continue predicting over many days ahead. Next in figure 15 is shown 2 years ahead prediction, surprisingly the model predicted the drop in price from covid in 2020 (with a small 1 month lag) from the drop in energy demand and the slow recovery from the lifting of the covid measures .

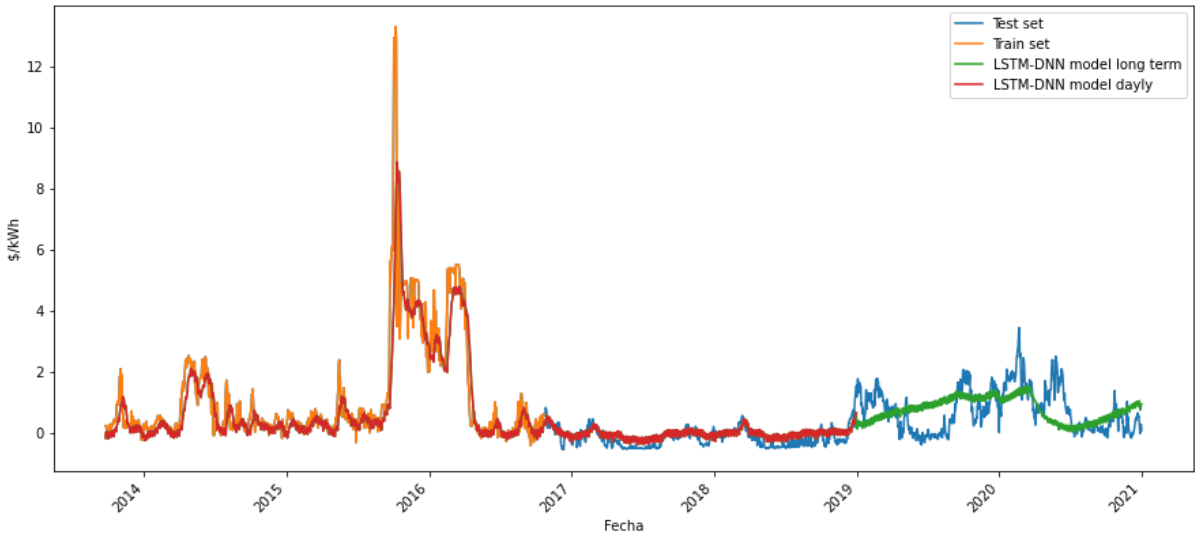


Figure 15. *Plot of part of train set (orange), test set (blue), daily forecast for LSTM-DNN model (red) and long term forecasting for LSTM-DNN (green).*

This model is way better than the ARIMA at long term forecasting, it seems that adding the correct energy demand solves the problem of getting stuck on predicting constant values. Also by taking annual averages (figure 16) we can see that our model does manage to stay close to the real value even after a 4 year prediction which is remarkable since our model was not trained or directly knows anything about the covid neither on the weather features on those dates..

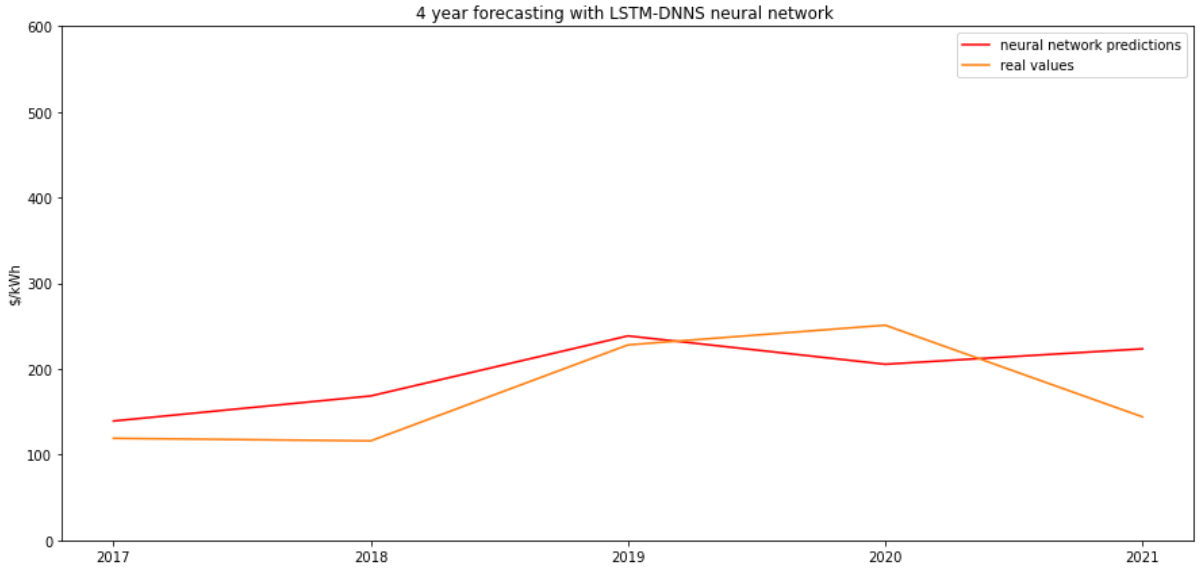


Figure 16. *Plot of 4 years term forecasting for LSTM-DNN (blue) and real averages from XM (orange).*

However we now need to not only predict the energy price but also the energy demand, this increases the difficulty of the problem but the values of the energy demand follows a monotonous behavior that makes it easy to reproduce making the model useful and scalable.

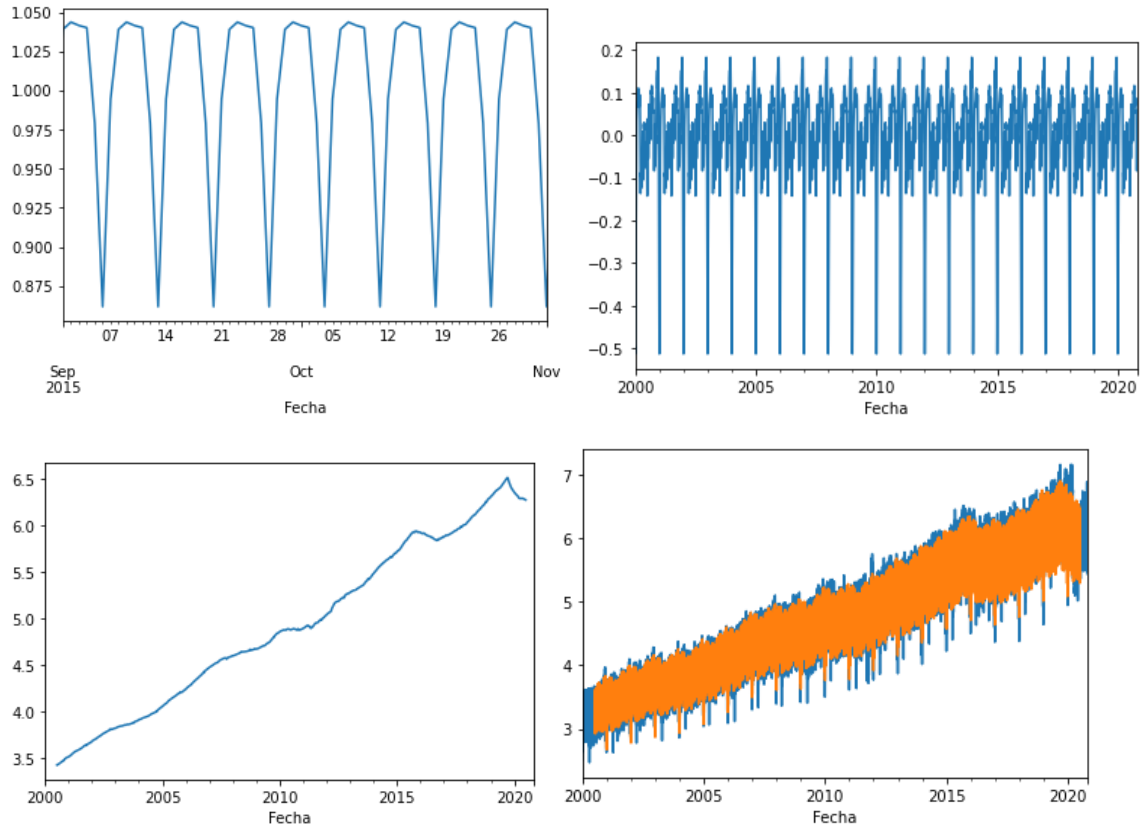


Figure 17. *Top left: weekly periodic behavior. Top right : yearly periodic behavior. Bottom Left: overall trend . Bottom right: real energy demand (blue) and predicted energy demand (orange)* .

From seasonal decompose analysis (figure 17) we see that there is a weekly multiplicative periodic part and a yearly additive periodic part, then the model follows an overall linear growth. By adding the linear part with the yearly trend and then multiplying with the weekly trend we can recover the energy demand with a sMAPE value of 2,8.

So by estimating the current overall slope we can quite easily forecast the energy demand for future years assuming that nothing big happens in that time (like COVID in 2020) and even if something like this happened this model can be easily fixed by taking new estimations on the overall slope from new data.

This model can be greatly improved by adding other variables (in this case we added the enzo values, the weighted average price for hydroelectric and thermal energy) alongside the energy demand as can be seen from figure 18.

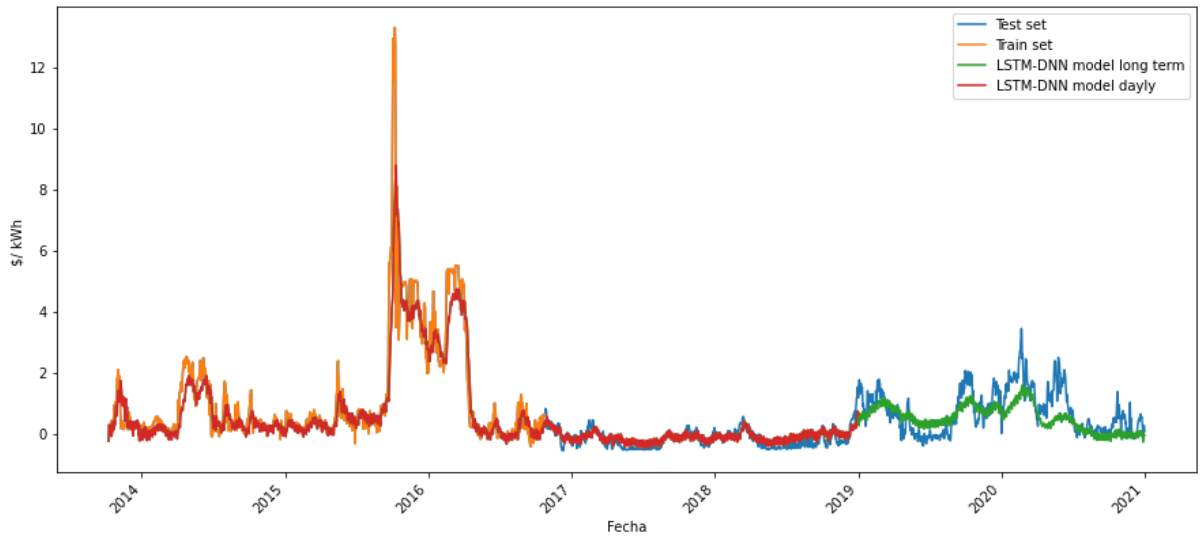


Figure 18. *Plot of part of train set (orange), test set (blue), daily forecast for LSTM-DNN model with more variables (red) and long term forecasting for LSTM-DNN with more variables (green).*

However we would need to have models to predict those variables to a certain degree of precision which is really difficult for variables like enzo is a really complex variable from which there is not yet an accurate method and investigation is still going trying to find such method, let alone the hydroelectric generation which depend on enzo and other weather variables.

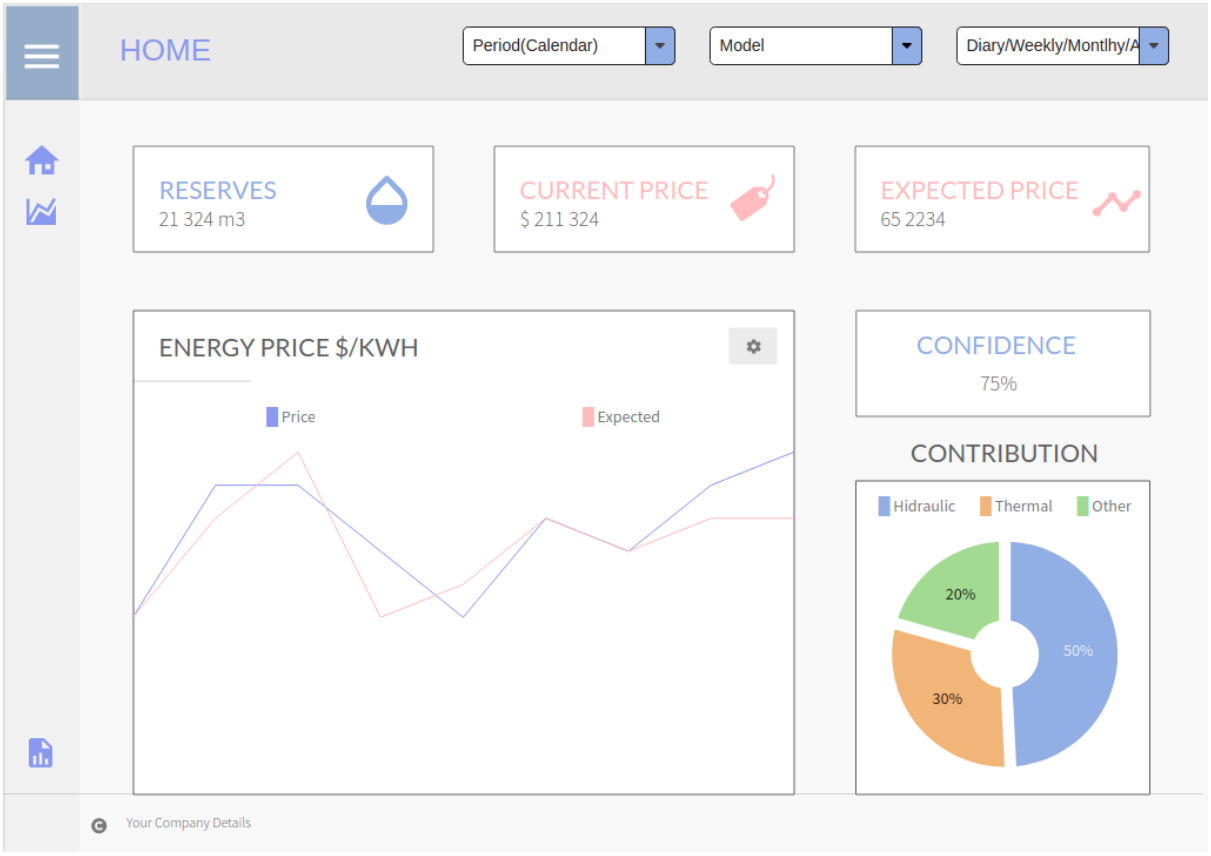
Mockup frontend design

Graphical interface aims to be a practical tool for Conexalab users. Hence, it has to be as intuitive as possible, and the more user-friendly the better. Keeping in mind that the main goal is to reduce the rate of return on investment on self-generation and distributed generation projects, it should be structured for solving that request.

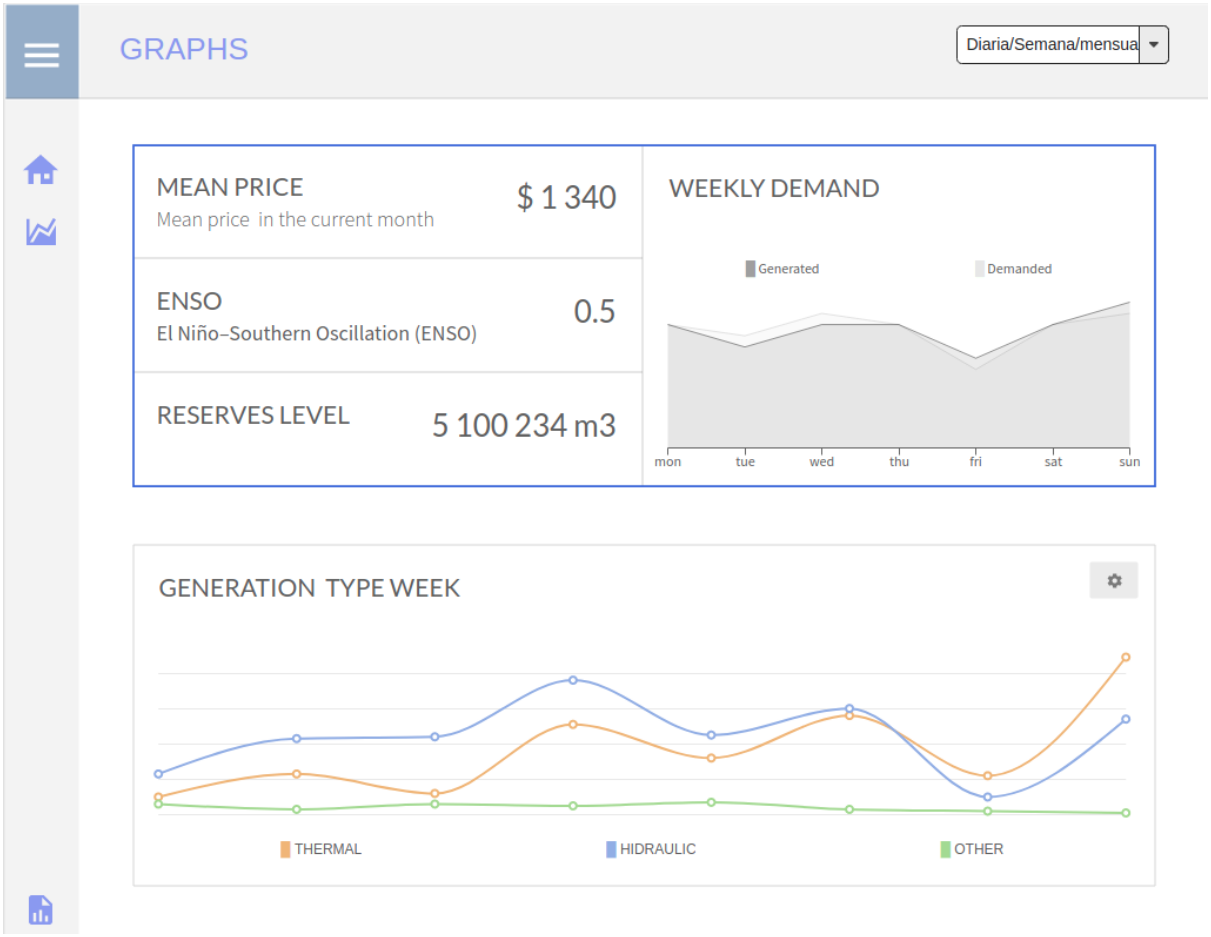
First of all, the user will have a Home Page with a summary of the principal variables included in the process. These values would show the state of the Colombian energy market at a given time.

Then, it is expected to have a how-to-use section, where the user will learn to use the different options and to flow through the colombian-energy-price-prediction algorithm. Plus, this component would explain the way to make a simple EDA (Exploratory Data Analysis) and get some conclusions based on the graphs.

Next, the user will have a platform for consulting some data related to energy price. It would have some useful graphs, for the sake of understanding the market behaviour.



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit tincidunt id. Sed rhoncus, tortor sed eleifend tristique, tortor mauris molestie elit, et lacinia ipsum quam nec dui. Quisque nec mauris sit amet elit iaculis pretium sit amet quis magna. Aenean velit odio, elementum in tempus ut, vehicula eu diam. Pellentesque rhoncus aliquam mattis. Ut vulputate eros sed felis sodales nec vulputate justo hendrerit. Vivamus varius pretium ligula, a aliquam odio euismod sit amet. Quisque laoreet sem sit amet orci ullamcorper at ultricies metus viverra.



Bibliography

[1] Hyndman. R , Athansopoulos. G **Forecasting: Principles and practice.** 2nd. Edition Available at https://otexts.com/fpp2/toolbox.html

[2] Weron, R. (2014). Electricity price forecasting: **A review of the state-of-the-art with a look into the future.** International Journal of Forecasting, 30(4), 1030–1081

[3] J. Lago, F. De Ridder, P. Vrancx, B. De Schutter **Forecasting day-ahead electricity prices in Europe: the importance of considering market integration** Appl Energy, 211 (2018), pp. 890-903, 10.1016/j.apenergy.2017.11.098

[4] S. Makridakis **Accuracy measures: theoretical and practical concerns** Int J Forecast, 9 (4) (1993), pp. 527-529, 10.1016/0169-2070(93)90079-3

[5] Jesus Lago, Fjo De Ridder, Bart De Schutter **Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms** Applied Energy Volume 221, 1 July 2018, Pages 386-405**, 10.1016/j.apenergy.2018.02.069**