Prediction of Energy Consumption in the Greater Buenos Aires Area

Career: Computer Engineering
Thesis Title: Master in Data Intelligence oriented to Big Data
Author: Franco Ariel Demare
Supervisor/Co-supervisor: Prof. Dr. Aurelio F. Bariviera
Author's Contact Email: frandemare@gmail.com
github:
https://github.com/francodemare/Energy-consumption-forescanting-in-the-Greater-Buenos-Aires-area/tree/main

Abbreviated Academic/Professional Biography

I began my studies in Computer Engineering in 2011 at the National University of La Plata and graduated in 2017. Since completing my degree, I have been working in IT companies, specifically in business process automation with various low-code software, Python, and the latest AI trends.

Keywords: *Machine learning; Statistical model; Energy consumption; Neural networks; Time series; Prediction*

Motivation

The primary motivation for this work lies in the increasing need to accurately predict electricity demand due to rapid economic expansion and the rise in energy consumption, both industrial and domestic. As societies and economies develop, energy demand grows exponentially, posing the challenge of ensuring an efficient and sustainable supply of electricity. Accurate consumption prediction is fundamental to optimizing resource management, reducing costs, and minimizing environmental impact, aligning with the Sustainable Development Goals (SDGs).

This thesis focuses on applying and comparing predictive models, both statistical and machine learning, with the goal of identifying the model that offers the lowest error rate in daily electricity consumption prediction in the Metropolitan Area of Buenos Aires.

Contributions of the Thesis

The central contribution of the thesis resides in the comparison and development of advanced predictive models for estimating electricity consumption in the Greater Buenos Aires (GBA) region, using an innovative approach with statistical and neural network models. The main objective of the work is to select and propose the most suitable model, based on the lowest prediction error and its capacity to be computationally scaled in data size.

The work introduces the implementation of traditional models such as SARIMA, known for its use in time series with seasonality, and more advanced models such as Support Vector Regression (SVM), with competitive results in terms of accuracy and processing times. Subsequently, more complex models based on neural networks such as CNN and LSTM are explored, offering notable improvements in energy consumption prediction.

The most significant value of the work is the introduction of a hybrid methodology, combining neural networks with Empirical Mode Decomposition (EMD) techniques, which allows for a considerable improvement in prediction accuracy. In particular, the LSTM-EMD model is consolidated as the best, with a MAPE (Mean Absolute Percentage Error) of 4.1%, demonstrating its ability to predict electricity consumption with high precision.

The thesis also highlights the focus on scalability and processing time, crucial variables for the implementation of these models in real-world environments with large volumes of data. The validation methodology used—a validation grid—provides stability to the results, allowing greater confidence in the evaluation of the best model. This less traditional approach avoids the simple separation of training and test data, ensuring more robust results.

In conclusion, this work not only advances the field of energy consumption prediction through machine learning techniques, but also proposes a scalable model, efficient in time and computational cost, suitable for future applications in other geographical contexts.

Data

The dataset used can be found on the website of the Secretary of Energy (http://datos.energia.gob.ar/dataset).

Within the site, a filter was applied by the field "Energy Demand" and the dataset called "Base Demanda Diaria 2017 a 2023.xlsx" was downloaded. The dataset can basically be seen as different time series.

These series contain, per day, the energy consumption expressed in MW (megawatt) of different regions of the country. In this work, only the values of one time series, the consumption of the Greater Buenos Aires region, will be used.

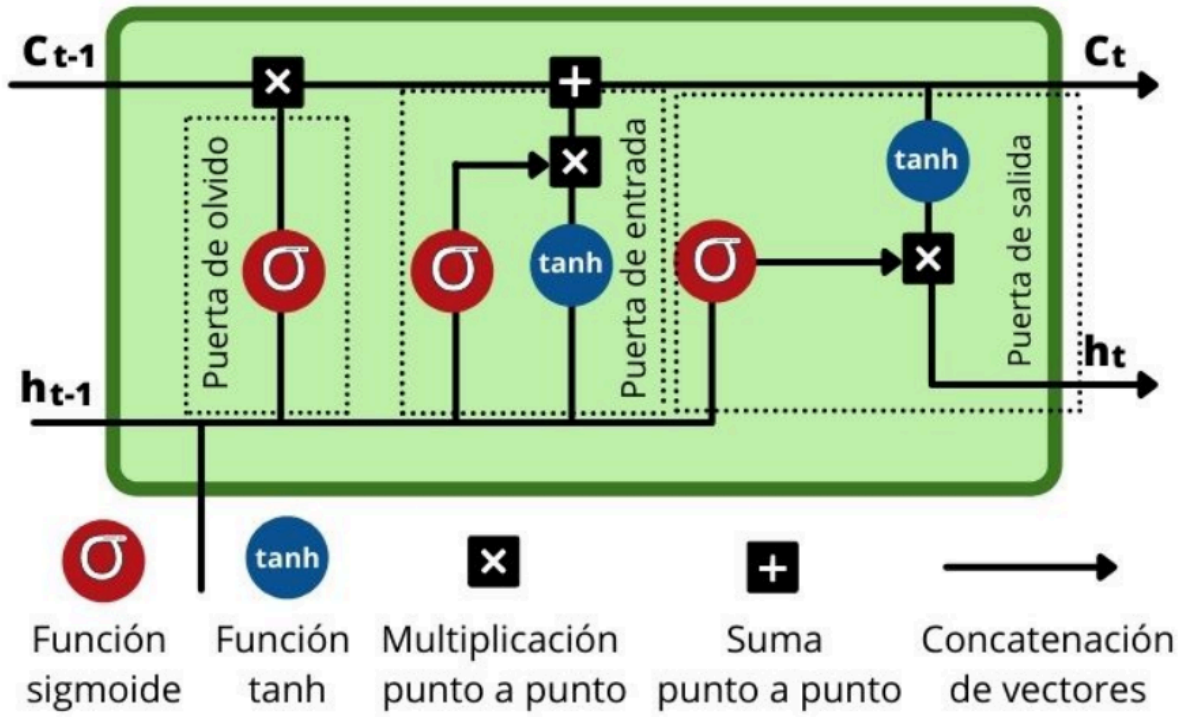
Stationarity

To verify the stationarity of the data, an Augmented Dickey-Fuller (ADF) test was performed on a specific dataset. The ADF test is used to statistically detect the presence of stochastic trend behavior in the time series of the variables. In this test, the null hypothesis is that the time series contains a unit root and, therefore, is not stationary. On the other hand, the alternative hypothesis is that there is no unit root in the time series. Therefore, to conclude that a time series is stationary, the null hypothesis must be rejected

Prediction Models

Long Short-Term Memory (LSTM) Neural Network Model

LSTMs are a type of recurrent neural network where each memory cell has a group of very specific operations that allow controlling the flow of information. These operations, called gates, allow deciding if certain information is remembered or forgotten. Within the memory cell, new information is added to that coming from previous sequences, that is, from previous time steps. The relevant new information is added to the flow thanks to an addition operation.



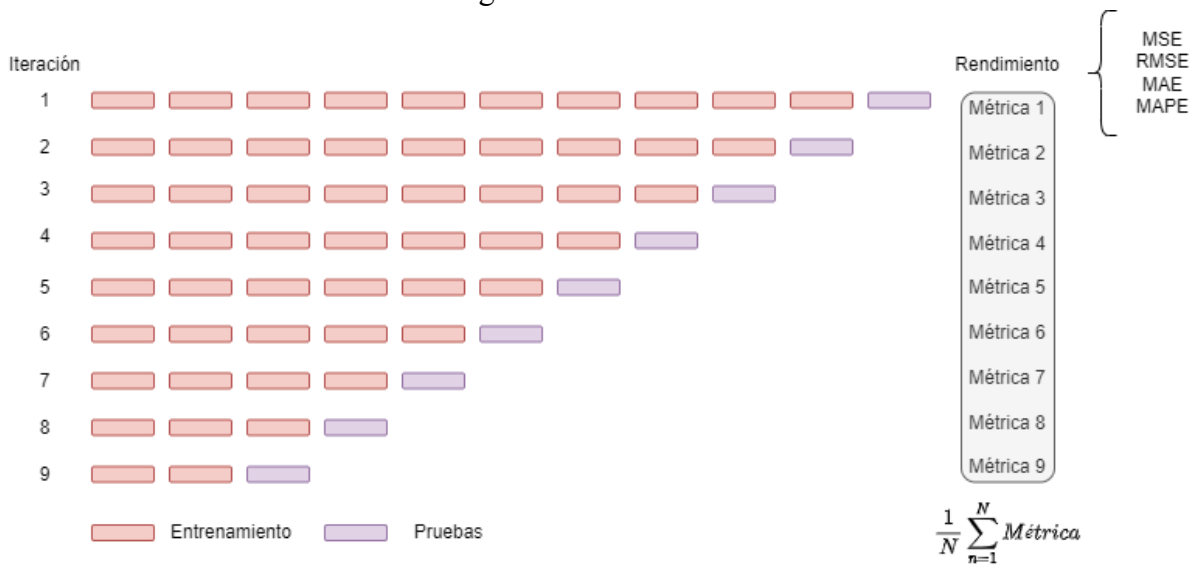
Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is an adaptive multi-resolution data technique to decompose a signal into physically significant components. The peculiarity of this technique is that it does not use any prescribed function, but adapts automatically based on the analyzed signal, hence the adjective "empirical".

In this case, it was decided to apply two variations of the Empirical Mode Decomposition: the original EMD algorithm and CEEMDAN. For this, the EMD library was used.

Validation Grid

A batch validation of the models was generated.



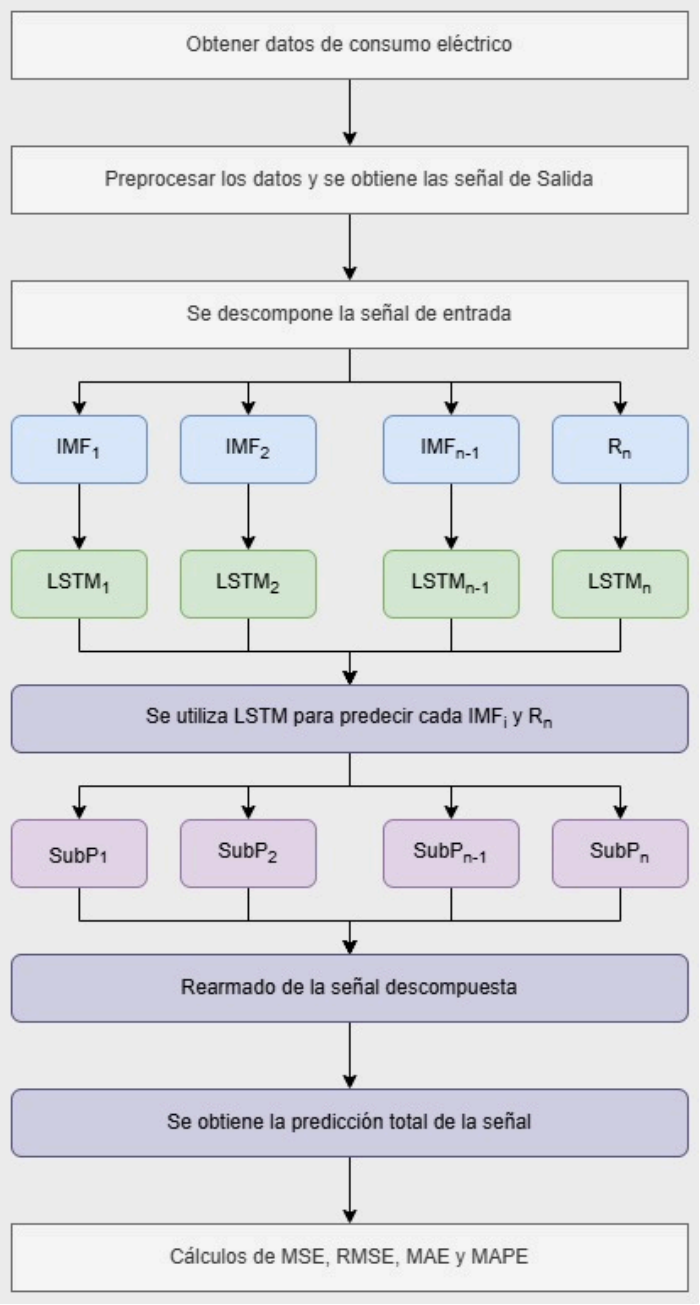
Each model was evaluated using this same type of validation. The method consists of taking an initial batch of data, in this case, it began by taking all the data. In this first stage, the batch is separated into 80% training data and 20% test data. The model is executed and then the four error measurements detailed above are obtained.

Then, in the next iteration, another batch is taken but now with a reduced set of data, for this specific case, a reduction of 20% of the initial batch was made. The model is executed again, separating the remaining data into 80% for training data and 20% for test data, and the error measurements are taken again. In this way, the error measurements are accumulated in variables and the batches, for each iteration, are reduced. These steps will be repeated until the batch reaches a minimum data size. For this development, the minimum size is 15% of the dataset.

Final Architecture of the Proposed Model

As a first instance, the electricity consumption data is obtained from the downloaded dataset, after processing these data, as mentioned in the section.

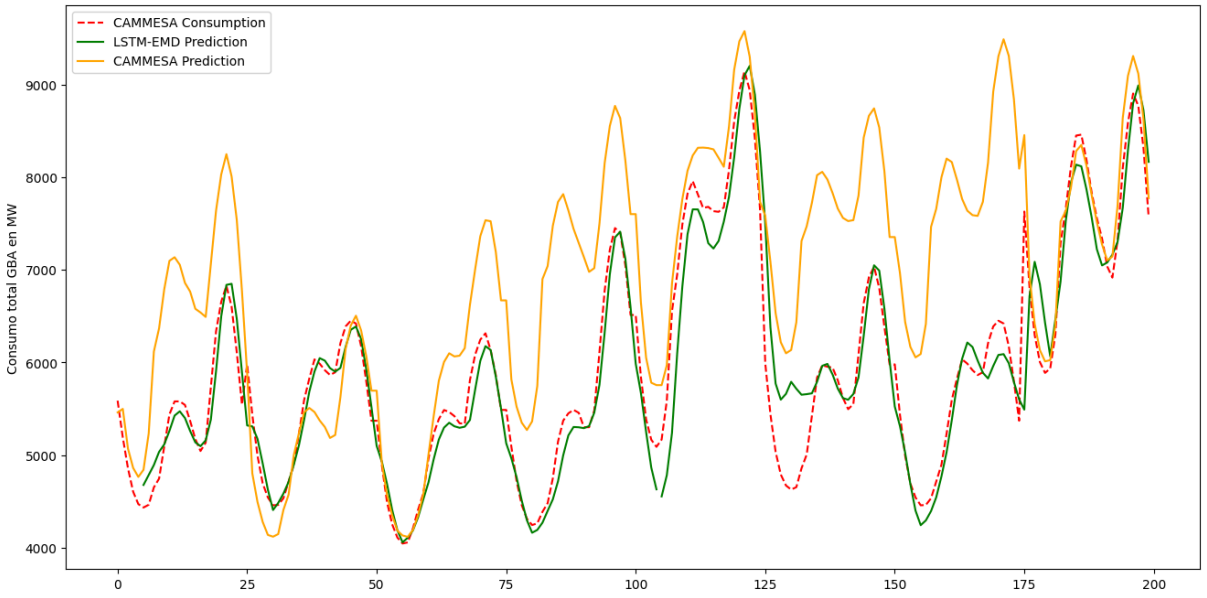
Then, the Empirical Mode Decomposition is applied, which gives us a series of new generated functions, to then apply the LSTM model to each of them. Finally, the predictions of each of the functions are reassembled to obtain the final function and prediction. As an additional step, the calculations of the errors defined above are performed to check the model's performance.



Model Prediction

Official intraday data from CAMMESA (electricity market regulator) were obtained. With these intraday data, the calculation of the errors between the real consumption and the prediction made by CAMMESA was performed: RMSE 1336.17; MAE 1044.9; MAPE 14%.

These errors will serve as a comparison against the model implemented in this thesis. The prediction was made with the LSTM-EMD model, where only 2 IMF functions were used. The results of the errors in our experiment were: RMSE 501; MAE 182; MAPE 3.1%. The processing time was 37 seconds.



In the previous figure, the discontinuous red line can be seen representing the real consumption per hour for each of the 201 hourly observations, in green the prediction made by the LSTM-EMD model, and finally in orange the consumption predicted by CAMMESA. The latter, it can be seen that although it accompanies the seasonality of the red curve, tends to give consumption results much higher than those that result finally.

Comparing the errors reported by CAMMESA and that of our model, it can be observed that the prediction errors of CAMMESA are between 3 and 4 times higher than the LSTM-EMD model. Although information is not available about how the predictive model is implemented by CAMMESA, it can be inferred that it performs a statistical prediction of the data. A possible explanation for the consumption overestimation error may be due to the company's own function: to plan the energy capacity needs, coordinate the dispatch operations, and regulate the economic transactions of the wholesale electricity market. This means that, to guarantee sufficient production to cover consumption, it prefers to overestimate the predictions. As a consequence, it induces a higher production to avoid an eventual lack of supply.

It is justified to run the model simply with two IMF functions because it is the minimum number of generated functions, with which good results are obtained. The prediction can be improved by generating the optimal number of IMF functions, but the computational cost does not justify the prediction result.