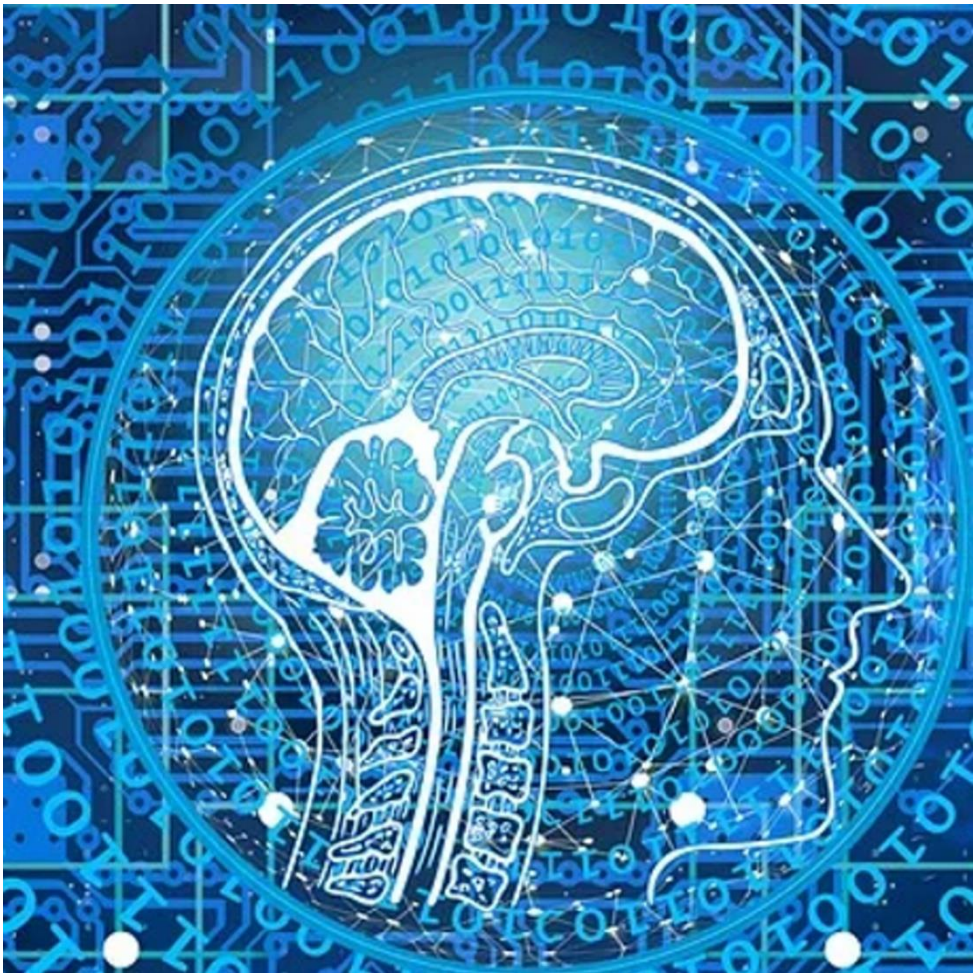
Specialized Models: Time Series and Survival Analysis

Final Project



Gabriel Moraes Magalhães

Summary:

Main Objective of The Analysis:

d

Brazil is one of the biggest generators of renewable energy, the main source being hydroelectric. So, the main objective of the analysis is to predict the national energy demand for each month in a year ahead, to ensure that we'll be able to supply the total demand. For this I'll be using a ARIMA model with seasonality.

The Dataset:

The dataset that I'm using came from Kaggle, and is a time series of the energy demand each hour in Brazil of the last 23 years. The original source of the data is ONS, the main organization responsible for the energy management in the country.

The dataset has only two columns:

- index: Timestamp of the hourly demand
- hourly demand: The energy demand in MW

Data Exploration and Cleaning:

Some actions that I did in the EDA and cleaning step:

- Simply plotting the complete series, I could see that the data has an outlier, and replace the value.
- The data from before 2003, has some anomality, so I discarded the tree firsts years of data from the dataset.
- I did a down sample in the data from hourly to monthly, because monthly data is more adequate to the business.
- I did a multi-line plot of the years to visualize if the dataset would have some evident seasonality, and it has.
- I plotted histograms of the series and the series with one differentiation, for the hourly, daily, and monthly series to check if we have a normal distribution in the data.
- I runed an Augmented Dickey-Fuller test, to see if the series was stationary, even though the series having a clear trend and seasonality.
- I plotted ACF and PACF plots for the hourly, daily, and monthly series.
- I applied a triple exponential smoothing to make the data less noisy.

Model Training and Selection:

For model training I try to reach the ARIMA parameters by running some ACF and PACF plots, and testing some differentiations. The models that I ended with this "manual" approach are:

- ARIMA (1, 1, 2) x (1, 1, 2, 12) (which is the manual model with the best scores)
- ARIMA (2, 2, 3) x (1, 2, 2, 12)

I have also used pmdarima to brute force model testing and selection, and the best model that this automatic method reached is:

- ARIMA (1, 1, 0) x (1, 1, [1], 12)

The pmdarima model give me sightly better results, so I ended with it.

Model Recommendation:

The model that I recommend in this scenario is the same that I ended with, because ARIMA with a seasonal component works really well when we have a good amount of data, like the twenty years that I used to train the model, and also works well when the process that generate the data is the same over the time, which is my case after I removed the years before 2004.

Furthermore, ARIMA-based models are easy to implement and explain than Neural Networks-based models, which also would works great in this case after some preparations.

Findings and Insights:

- The energy demand keeps rising, as expected.
- We have a clear seasonality.
- The period of November to March is when we have the higher energy demand and is when we have the summer season here in Brazil. So, the higher energy demand could be caused by the intense use of air-conditioners in hot periods.

Next Steps:

- Study more about neural networks and apply to this case to see if I can get better results.
- Analyze the inconsistent behavior for the years before 2004, and see if this additional data can improve the model.
- Analyze if the covid-19 pandemic causes any impact in the energy demand.
- In the future, compare the 2023 real energy demand with the predicted values.