BrainStation Data Science Capstone Project – Simón Román O. June 2023 – London (UK)

Introduction – This report aims to summarize and present the findings of the monthly time series forecast of UK's energy consumption. The business relevance of this project lies in the need for the different stakeholders of the energy sector to accurately predict the energy demand given how difficult and expensive is energy to be stored. Energy production is a catch game between demand and generation that moves as fast as the time it takes you to turn on the light switch.

The report will explain i) the source and characteristics of the data used; ii) the main discoveries of the EDA process and data wrangling; iii) models used to forecast; iv) results; and v) the next steps.

Methodology – The time series analysis was performed using statistical models and included data collection and exploration, parameter tuning and selection, model fitting, forecasting, and validation. The error metrics used to evaluate the models were the mean absolute percentage error (MAPE) and the mean absolute error (MAE).

Data source and characteristics – The data used for the time series was sourced from the National Grid webpage and included, in general terms, energy demand, energy import and export flows, and wind and solar capacity and generation. The data consisted of 48 daily measurements in 30-minute intervals and ranged from January 1, 2009, until January 1, 2023.

The target variable of the analysis is the National Demand which is "the sum of metered generation but excludes generation required to meet station load, pump storage pumping, and interconnector exports"[1] and is measured in MW. As the analysis seek to predict monthly energy consumption the data was resampled to a monthly frequency, which means that every month included the sum of all the data points recorded for all days in that month.

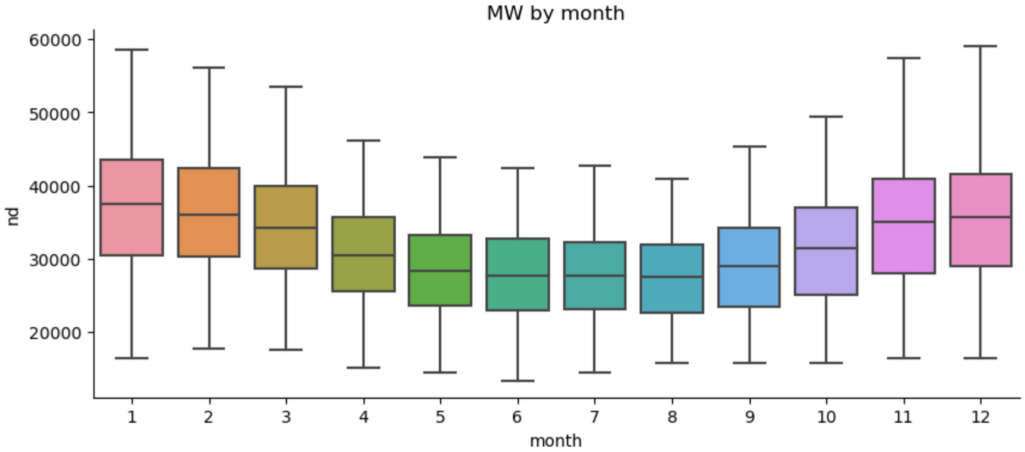
We regret that the model had so little number of data points, this meant that the model has less information to learn from and capture the underlying patterns or trends in the time series.

Exploratory Data Analysis (EDA) and Data Wrangling – The data had some issues regarding missing values which meant that some imputations had to be made. However, the number of values imputed was not significant. The dataset had no duplicated or null values. The dataset had seven columns which presented all the registries of the energy imports and export flows from the UK, since no one flow was of interest we chose to merge all the values in a single feature column.

During the EDA we also decomposed the series into the trend, seasonal and residual components and found: i) that energy National Demand has been decreasing steadily since 2013 with some small increase bumps; ii) our target variable clearly has a seasonal pattern that goes with weather seasons, high demand in the cold months of winter and low energy consumption in summer; and iii) that the residual component seems random, variation in this element is not high except for several steep high and low peaks.

The following plot shows the monthly decomposition of the data, it illustrates how energy consumption through the year has a clear seasonal pattern with low consumption in the warmer months and high consumption in the cold winter months.

[1] https://data.nationalgrideso.com/demand/historic-demand-data/r/historic_demand_data_2022



Baseline Models – For baseline models, we used a Simple Mean approach by calculating the mean of the ND column. In other words, for this model, our best estimate at any given point in time is that the temperature is equal to the average temperature of the entire variable. The other baseline model involves calculating the rolling average of the three previous periods. This method smoothed the data, resulting in reduced abrupt changes. The Simple Mean model had a test MAPE of 5.48% while the Rolling Average score was 5.13%.

Statistical Models – Three different statistical models were used to forecast the series: ARIMA, SARIMA, and SARIMAX. We will explain briefly what their main characteristics are, how we selected the parameter values, and what results we got.

ARIMA (Autoregressive Integrated Moving Average) - is a time series forecasting model that combines autoregressive (AR), differencing (I), and moving average (MA) components. This model is very useful to capture linear relationships, it presupposes that the series being forecasted is stationary. The ARIMA model has three parameters: p, d, and q. The 'p' parameter represents the number of lagged observations used to predict the current value in the time series. The 'd' parameter represents the number of times the time series needs to be differenced to achieve stationarity. And the 'q' parameter represents the number of lagged forecast errors used to predict the current value.

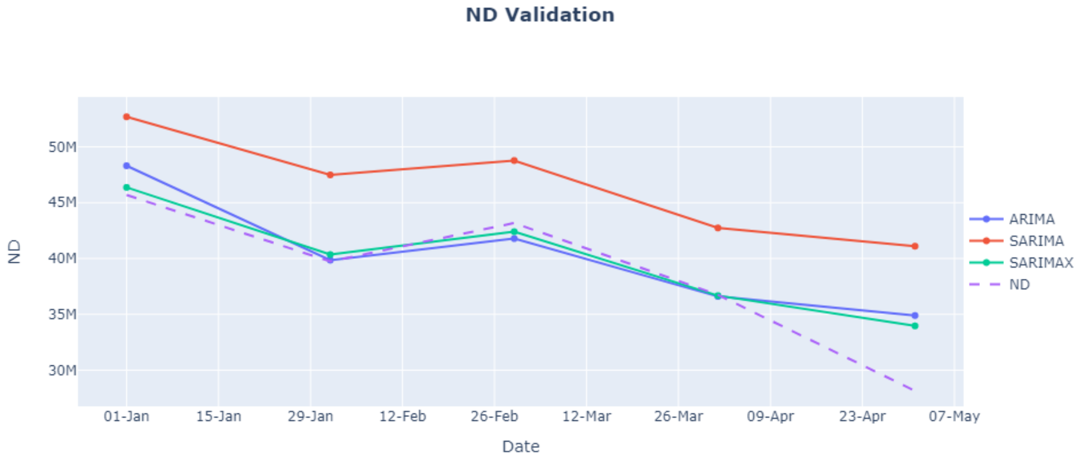
For our modeling, we chose the p, d, and q parameters by iterating over a function that looked for the least MAPE, to inform our results we also used the number of p-values with statistical significance. We opted for a (4, 0, 2) configuration, the zero for the d parameter responded to the use of differenced data which in turn resulted in a 5.15% MAPE in the test set.

SARIMA (Seasonal ARIMA) – is an extension of the ARIMA model with the added feature of accounting for seasonality in the data. The SARIMA model uses the same parameter as ARIMA plus the P, D, and Q parameters that inform the seasonal components. They are similar to the (p, d, q) parameters but specifically model the seasonal patterns of the data. The 's' parameter denotes the length of the seasonal cycle, indicating the number of time steps that define one complete season.

To choose the SARIMA parameters we ran a function that iterated on both sets of parameters and looked for the lowest Akaike Information Criterion (AIC). The chosen parameter configuration was (3, 1, 2)x(0, 1, 3, 12), with almost all of the coefficients having statistical significance. We had to fit this model in two different train and test datasets because at first it was not able to grasp the seasonal component, with the second train and test set the model had a MAPE of 7.05% in the test set.

SARIMAX (Seasonal ARIMA with Exogenous Variables) – this model is an enhancement of the SARIMA model that incorporates exogenous variables, which are external factors that can impact the time series. The additional parameter for this model is the exogenous variables. We ran the model using the same parameter as the SARIMA model and as exogenous variables initially, we chose import-export, embedded solar generation, and temperature. We ran a function that looped through the different iterations of the three exogenous variables and found that the better-performing variables were temperature and import-export. SARIMAX was our best-performing model, with a 3.73% MAPE score on the test set.

Results – To further evaluate the model's results we contrasted them with new data from 2023 ranging from January to May. This unseen data gave us insights on the model forecasts and the results were not very good. The next plot illustrates how the three models had similar behavior (SARIMA was way off) which emulated the actual data but missed the continuing downward trend in energy consumption in April.



All the models increased their MAPE regarding the validation set. The results are presented in the following table:

Model	Train MAPE	Test MAPE	Train MAE	Test MAE	Validation MAPE	Validation MAE
Simple_Mean	3.95%	5.48%	1,867,551	2,126,466	-	-
Rolling_AVG	5.31%	5.13%	2,520,044	2,122,838	-	-
ARIMA	12.24%	5.15%	6,198,868	1,990,700	6.71%	2.202,055
SARIMA	4.56%	7.04%	2,184,182	2,740,519	22.02%	7,859,852
SARIMAX	3.43%	3.73%	1,608,039	1,417,534	5.13%	1,591,380

Overall, the results might indicate that the models tended to overfit, which means that the model captured noise or random fluctuations in the data instead of the true underlying patterns. This might have been caused by the small number of datapoint used to fit the model.

Next Steps – To achieve a better performing model, in further iterations, we would aim to: 1. improve our results from the SARIMA model (which in turn informed the SARIMAX model) which theoretically should be better than those of the ARIMA because of the seasonal behavior of energy consumption; 2. Inform the SARIMAX model with more exogenous variables in order to improve performance; and 3. Test machine learning algorithms and contrast them to our statistical models.