Laurence Forget Brisson

Capstone three – Final Report

Presented to M. Ricardo D. Alanis-Tamez

Problem statement

Can we predict the electricity load of Panama, considering meteorologic conditions, holidays, and school schedule of three major cities in the country?

Context

Electricity needs in each country are evolving along, with an increased pressure from multiple components, including demographic growth, industrialisation, and transition towards electric resources. It is becoming increasingly important to ensure the integrity of the electricity distribution system to avoid any overload or disturbance, and to foresee the need of infrastructure development. Efficient electricity load forecast is a relevant avenue to deal with this problem.

Datasets

To address this question, a dataset was obtained from *Kaggle*, available as part of a competition named *Electricity Load Forecasting*[1]. This dataset contains the total electricity load of the country (in GW.h), as well as four climatic indicators (temperature (°C), relative humidity (%), liquid precipitation (mm), and wind speed (m/s)), provided for three major cities of Panama (Tocumen, Santiago and David). All variables are recorded hourly for the period between January 3rd 2015 and June 27th 2020, for a total of 48 048 entries.

Approach

The selected approach was to perform time series analysis and modeling to forecast the electricity load target as is (univariate), but also considering the other features as exogenous features (multivariates). Three models were selected and used to forecast electricity load, which are:

1. ARIMA (AutoRegressive Integrated Moving Average);
2. VAR (Vector Autoregressive); and
3. Prophet, from Facebook.

Data wrangling

The dataset was already quite clean, that is without any missing values, duplicated data and superfluous features to drop. The following steps were achieved to perform data wrangling:

- Column names were changed to be more meaningful and intuitive;
- Temporal data were assigned to a datetime format and set as the index;
- All other features' data types were changed to float to facilitate modeling;
- All features were observed through time to see evolution or occurrence, depending on the case;
- The presence of outliers was verified, and no value was discarded.

[1] https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting

Exploratory data analysis

Correlation between features were explored on a yearly, monthly, weekly, daily and hourly basis in order to also capture the temporal subtleties in the data.

Here are some insights about the correlation exploration:

- The target value, electricity load, does not seem to have a significant correlation with any individual feature, except for the hourly basis where a relation can be perceived with the temperatures of all three cities (correlation factors are 0.92 for Tocumen, 0.92 for Santiago, and 0.91 for David City).
- There is general correlation between temperatures of all three cities, with variation on the different scale observed, the weakest being on the hourly basis. The situation is similar for the humidity level, precipitation, and wind speed features, as well as for some of these features between them.
- Relations between features are more complexes on an hourly basis, with a strong cyclicality.

The target feature (electricity load) was then resampled on a weekly basis for noise reduction and seasonal decomposition was used to visualize the different components of the time series (figure 1), which are trend, seasonality, and residual.

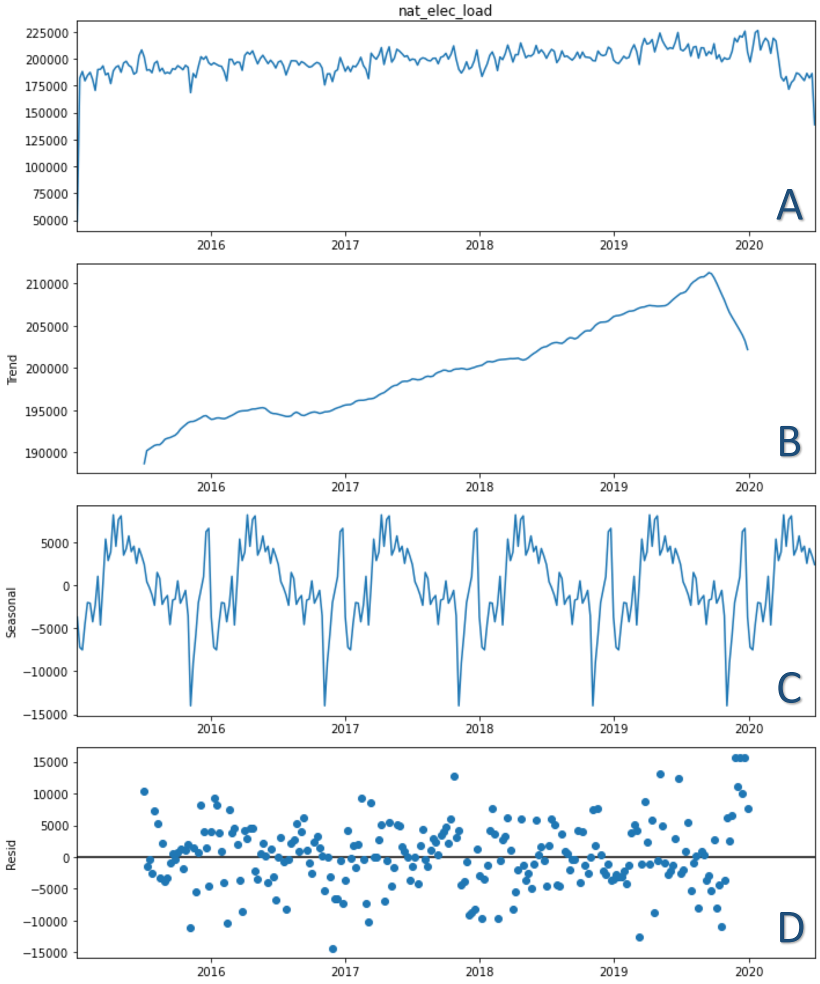

Figure 1 : Seasonal decomposition of the target feature (electricity load) resampled on a weekly basis. A: Original values; B: Trend; C: Seasonality; and D: Residual.

Seasonality of the target feature was further analysed, aiming to choose between ARIMA and S(easonal)ARIMA models. No clear pattern of seasonality was identified when using different level of differentiation (figure 2). However, a strong seasonal component can be noticed in figure 1C. This ambiguity will drive towards testing both models to verify the impact of this component on the forecasted results.

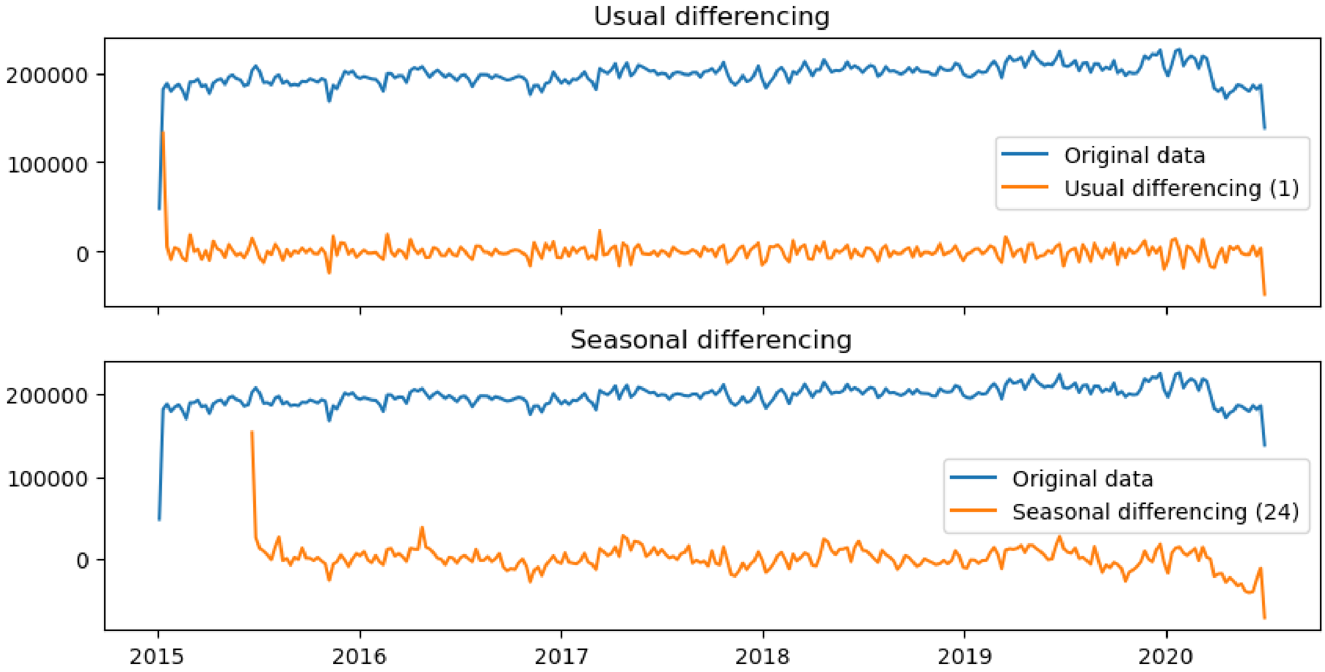

Figure 2 : Detection of a seasonal pattern using differencing. Up: One differencing; down: 24 differencing.

The occurrence of each category of the categorical features, which are holiday type and occurrences, as well as school days was then explored. There are 22 different holidays per year, but the occurrence of each is varying slightly throughout the year.

While 2020 is not recorded in its entirety, the beginning of this year is also affected by the global pandemic. This unexpected event had a great effect on the electricity load, as can be seen in both previous figures by a drastic drop in electricity load.

Pre-processing and training data development

Categorical features previously mentioned were leaved as is as they were already included as numerical data in the dataset. All 2020 data were dropped to avoid dealing with the effect of the global pandemic and to focus on complete years, from the beginning of 2015 to the end of 2019. Data was then resampled on a weekly basis using the sum or the mean, depending on the feature. Electricity load, precipitation, holiday, and school days were summed, while temperature, humidity and wind speed were averaged. First and last row were dropped to avoid dealing with incomplete weeks.

Data is considered as stationary when it's statistical properties (mean, standard variation and auto-correlation) are constant through time, and most time series models are designed to be applied on stationary data. Stationarity in the target data was verified using rolling mean and rolling standard deviation on a five-weeks window. A noticeable increase in the mean through time indicated a non-stationary behavior, and an augmented Dickey-

Fuller test was further used to assess on the stationarity status of all the features. The electricity load target feature is the only one having a p value over the threshold (p value = 0.16 > 0.05), validating the non-stationarity of the data. It was found that one-order differencing is sufficient to induce stationarity in the data (p value = 1.34e-22).

Autocorrelation and partial autocorrelation functions (ACF and PACF, respectively) were then used on undifferentiated (actual) and one-order differentiated (diff) data to understand the time series on hand and identify the best parameters for ARIMA modeling (figure 3).

ACF and PACF are used to identify how time series data are correlated with its own lags, including or excluding indirect correlations when solving the equation. In the case of the ACF, it has to be noted that actual data are highly correlated, and do not decay to zero, indicating again that the data is not stationary. When differenced, 2 branches stand out of the confidence interval (95%), while 7 are out in the PACF plot. This may indicate the presence of some pattern in the time series (*e.g.* seasonality), hence inducing some complexity for forecasting.

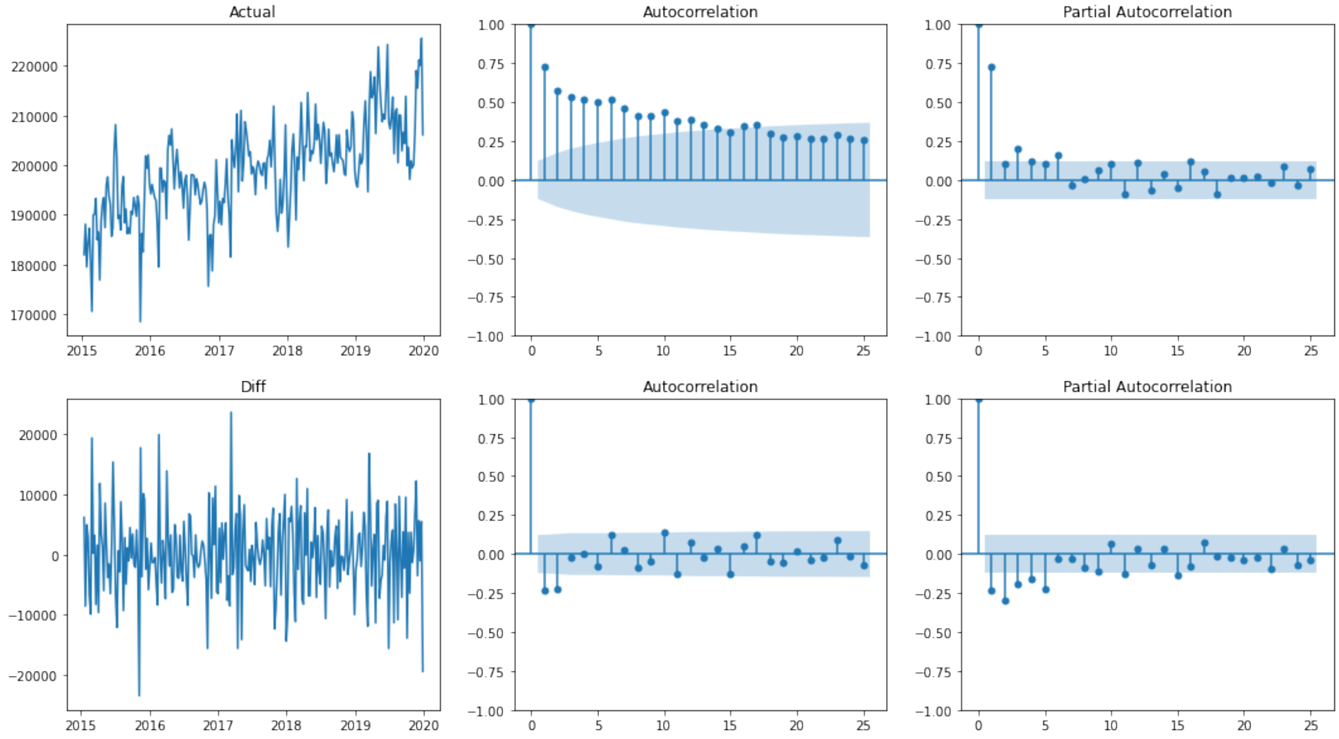

Figure 3 : ACF and PACF plot for the non-differentiated and one-order differentiated target feature (electricity load).

The dataset was finally split into training ang testing in order to train on 4 years (2015 to 2018) and test on 1 year (2019).

Modeling

Modeling is used in this project to forecast the electricity load through time. As stated earlier, three main time series models were selected for this aim, which are ARIMA, VAR and Prophet. All forecasts are shown individually in figure 4 and can be compared in figure 5.

ARIMA

This model decomposes into three terms, which are p (AR), referring to the number of lag observations, d (I), enclosing the degree of differencing, and q (MA), denoting order of the moving average, to which need to be added P, D, and Q and m (frequency of seasonality 0) if there is a seasonal component in the target feature and the model convert to a SARIMA. This model type was applied in 5 different ways:

1. Manually applied univariate ARIMA (figure 4A);
2. Manually applied univariate SARIMA (figure 4B);
3. Univariate Auto-ARIMA (figure 4C);
4. Multivariate Auto-ARIMA using other features as the exogenous variables (figure 4D);
5. Multivariate Auto-ARIMA using seasonal index as the exogenous variable (figure 4E).

In all cases, d and D parameters were set to 1 to indicate that one-order differentiation induce stationarity in the data.

In the manually applied univariate ARIMA (figure 4A), many parameter combinations were tested for p and q, but all provided a flat fit, indicating that the model failed to capture the details and specific components of the dataset.

A univariate SARIMA (figure 4B), also manually applied, was then tried to comprehend a seasonal component in the data, using the following parameters (p = 1, d = 1, q = 1, P = 0, D = 1, Q = 1 and m = 52). The provided fit is generally very good but tend be less accurate towards the end of the predicted period, indicating that a one-year period of prediction may be too long.

The focus was further turned towards auto-ARIMA. Univariate auto-ARIMA (figure 4C) indicated that the best parameter combination was (p = 0, d = 1, q = 0, P = 0, D = 1, Q = 0 and m = 52), hence validating the SARIMA approach previously attempted. The provided forecast seems to be less of a good fit with the testing set then for the manual SARIMA, leading to a generalized overestimation throughout the forecasted period.

Two multivariate auto-ARIMA were tested using the other features of the dataset (figure 4D), as well as the seasonal index (figure 4E) as the exogenous variables. Both versions of the model indicated the same parameter combination (p = 0, d = 1, q = 1, P = 0, D = 1, Q = 1 and m = 52) and have the exact same forecast. This may indicate that exogenous variables have a very negligeable impact in the model, or this could also imply a misapplication of the models. Moreover, the forecast provided by the manually applied univariate SARIMA (figure 4B) is also indistinguishable from both multivariate auto-ARIMA forecast (figure 4D-E).

VAR

This model is a generalization of the autoregressive model (AR) but incorporating a more dynamic approach through time using lags. The number of lags used in the model have a considerable impact on the forecast. This parameter is chosen by selecting the number of lags with the lowest Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Final Prediction Error (FPE) and Hannan–Quinn information criterion (HQIC). In this case, 14 was prescribed as the best number of lags. However, the forecast using 14 lags (figure 4F) provides a poor fit very prone to overfitting and overestimation of the cyclicity in the data. The model was then retried using a lower number of lags (10; figure 4G) and the result is much more satisfying, with a generally good fit which is smoother than the original data. Again, the fit is less accurate in the last portion of the forecasted period. The latter situation appears to be common to all models.

Prophet

This model uses additive regression, including linear or logistic growth to forecast. No extra parameter is required to run the model, which result in a very simple application. The forecast using Prophet (figure 4H) has a relatively good fit with the tested set, but with a small, generalized underestimation and what appears to be a temporal offset in the prediction.

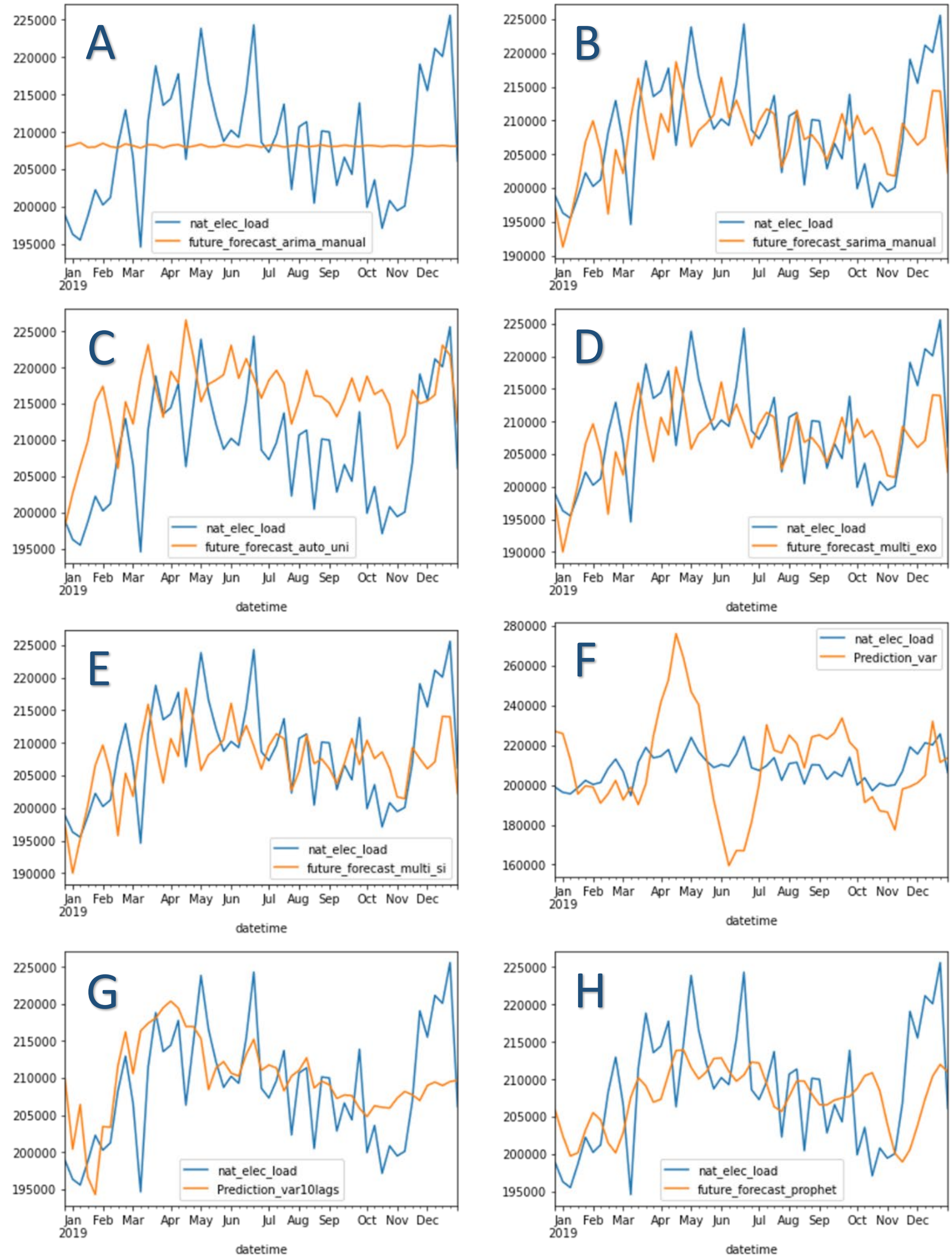

Figure 4 : Forecasted compared to actual data for 8 modeling (A: Manually applied univariate ARIMA; B: Manually applied univariate SARIMA; C: Univariate auto-ARIMA; D: Multivariate auto-ARIMA using other features; E: Multivariate auto-ARIMA using seasonal index; F: VAR with 14 lags; G: VAR with 10 lags; H: Prophet.

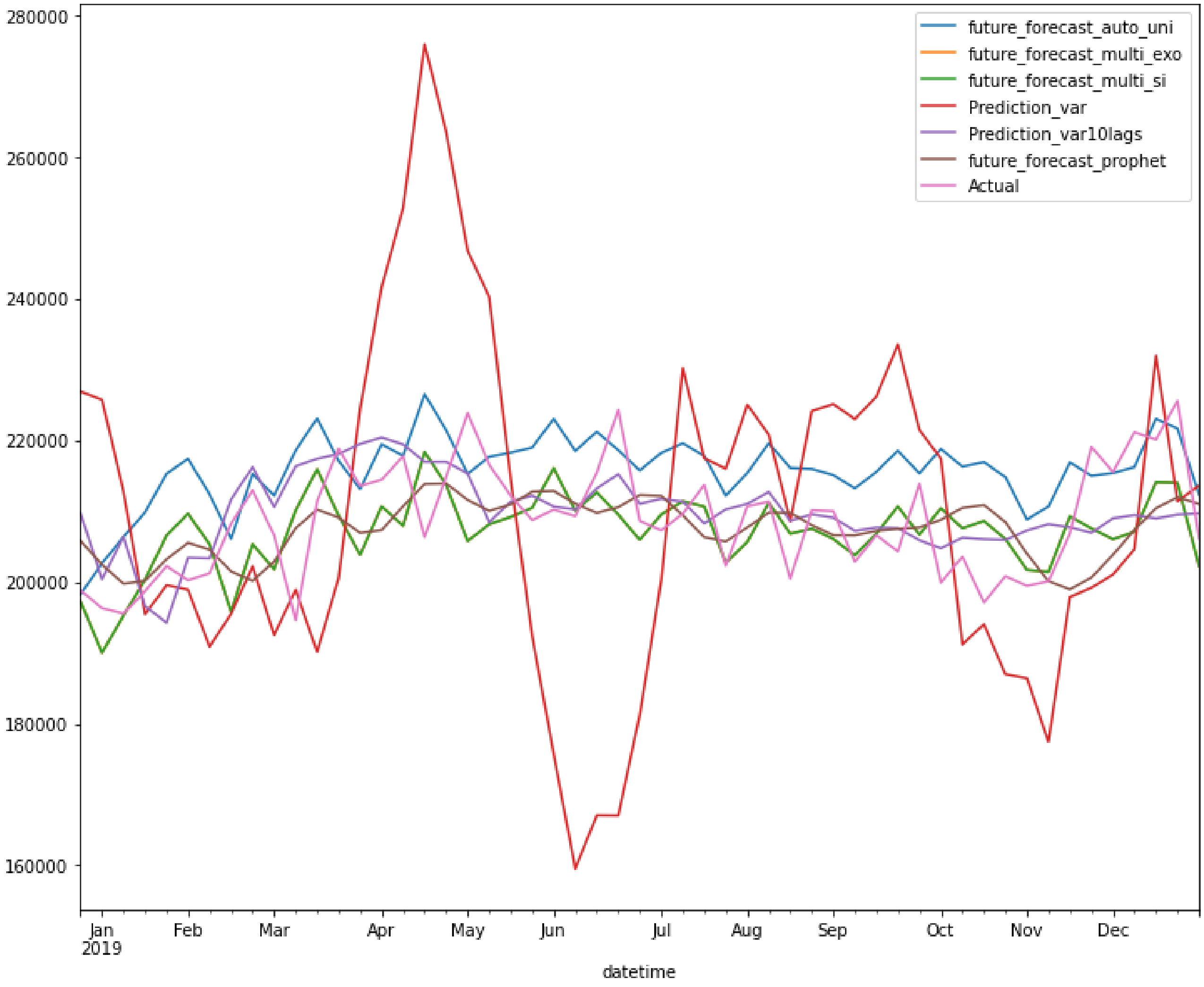

Figure 5 : Comparison of the forecast of all models with the test set (Actual).

Model review and recommendations

Between the different tested models (figure 5), both the manually applied univariate SARIMA and the VAR using 10 lags have the best results, according to most metrics (table 1) and to the visual fit of the forecasted values compared to the test set. By the observation of the ME metric, it can be noticed that the former model tends to slightly underestimate the test set, while the latter, at the opposite, tend to slightly overestimate it when forecasting. This particularity could be taken into account for model selection, given that over- or underestimation could be more useful or detrimental depending on a specific context of application.

Moreover, two advantages could be expressed in favor of the univariate manually applied SARIMA model. Firstly, it has a very transparent yet relatively simple application, either using manual parameter selection or when using auto-ARIMA. This can be viewed as an advantage over other models for customised implementation, as well as for results interpretation for a deeper interpretation of a specific dataset. Secondly, using only a single feature for univariate modeling may considerably reduce implementation time and resource.

It must be noted, however, that all models had a poor performance in forecasting the last quarter on the year. This may indicate that prediction should be performed on a shorter term and reviewed frequently to increase accuracy and general performance. Yet this could also suggest that the reviewed models may be less suitable to predict more abrupt changes in the data.

Table 1 : Model performance comparison

Metrics	Univariate ARIMA (Manual)	Univariate SARIMA (Manual)	Univariate auto-ARIMA	Multivariate auto-ARIMA (with other features)	Multivariate auto-ARIMA (with seasonal index)	VAR (14 lags)	VAR (10 lags)	Prophet
MAPE (%)	3.09	2.77	4.08	2.78	2.78	9.04	2.63	2.94
ME	-561	-1163	7155	-1504	-1504	1082	1239	-1159
MAE	6471	5831	8370	5858	5858	18991	5470	6172
MPE (%)	-0.13	-0.05	3.53	-0.62	-0.62	0.59	0.69	-0.44
RMSE	7819	7306	10014	7380	7380	23764	6985	7503

Where:

MAPE: Mean Absolute Percentage Error;
ME: Mean Error;
MAE: Mean Absolute Error;
MPE: Mean Percentage Error.
RMSE: Root Mean Squared Error.

Conclusion

To conclude, evolution of the electricity load of Panama can be successfully forecasted using multiple models. Due to transparency and simplicity of application, as well as the minimalism implied by using only the target feature, univariate SARIMA is recommended as the best model to use in this case. However, even the more accurate model would have failed to predict the effect on the 2020 global pandemic on the target feature.

Github links

Data wrangling

https://github.com/LaurenceFB/Capstone-3/blob/main/LForgetBrisson_Capstone3_DataWrangling.ipynb

Exploratory data analysis

https://github.com/LaurenceFB/Capstone-3/blob/main/LForgetBrisson_Capstone3_EDA.ipynb

Pre-processing and modeling

https://github.com/LaurenceFB/Capstone-3/blob/main/Capstone3_Pre-processing%20and%20modeling.ipynb

Metric file

https://github.com/LaurenceFB/Capstone-3/blob/main/capstone3_MetricsFile.csv