Electrical Distribution Network

Energy Consumption Forecasting

based upon

Victorian MRIM Meter Data

Khong Fwu Chin

18 September 2020

Abstract

This report documents the design and testing of machine learning models to forecast energy consumption for an electrical distribution network based on Victorian MRIM meter data. In addition, the report discusses where the data is acquired and the analysis of the data to ensure the data's viability as input for the machine learning models.

The performance of each model was compared with actual values based on visualisation and two performance metrics i.e. R2 Score and Mean Absolute Percentage Error (MAPE). The project has resulted in a satisfactory outcome where it provided recommendations on the most appropriate model for forecasting energy consumption for an electrical distribution network.

The report also documents additional work to explore further to this project. This includes the possible pipeline that can be used to forecast energy consumption for other electrical distribution networks. Besides energy consumption, this pipeline can assist in forecasting consumption of resources to generate energy.

Table of Contents

Background

Energy consumption forecasting plays a vital role in planning, operations and management of modern power systems. If the supply exceeds the demand, the energy can be stored. However, the surplus energy cannot be stored in a large capacity, has a limited storage lifespan, requires high maintenance and hence, the storage is costly. If the supply power goes below the demand, this would lead to overloading the supply line and causing potential blackouts.

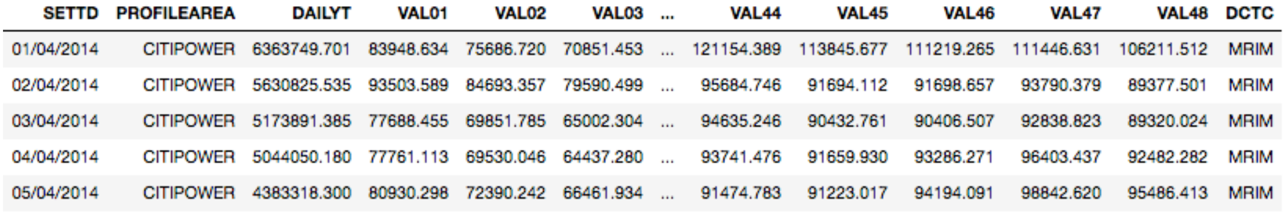
Objective

Providing reliable forecasting of energy consumption results in a better management of the electrical distribution network. The objective of this Project is therefore to forecast energy consumption for an electrical distribution network based on Victorian MRIM meter data. The forecast will be built upon machine learning models that is utilised in time series problems. In addition, the energy consumption will be forecasted based purely from historical data.

The energy consumption will be forecasted based on historical data that are obtained directly via MRIM meters from an electrical distribution network. The rationale behind this is to alleviate the dependence of other parameters, such as the weather, which are in itself also forecasts. If the weather or such similar parameters are forecasted incorrectly, it will directly impact the accuracy of the energy consumption forecast.

Data Source

The Australian Energy Market Operator (AEMO) are responsible for planning, developing and operating markets that respond to Australian gas and electrical needs. The data was thus acquired from the AEMO website for the particular electrical distribution network. The data contains the aggregated half hour energy consumption (VAL01~VAL48) and the daily total (DAILYT) and these data are released on a quarterly basis. A sample of the data is presented in Figure 1.



SETTD	PROFILEAREA	DAILYT	VAL01	VAL02	VAL03	...	VAL44	VAL45	VAL46	VAL47	VAL48	DCTC
01/04/2014	CITIPOWER	6363749.701	83948.634	75686.720	70851.453	...	121154.389	113845.677	111219.265	111446.631	106211.512	MRIM
02/04/2014	CITIPOWER	5630825.535	93503.589	84693.357	79590.499	...	95684.746	91694.112	91698.657	93790.379	89377.501	MRIM
03/04/2014	CITIPOWER	5173891.385	77688.455	69851.785	65002.304	...	94635.246	90432.761	90406.507	92838.823	89320.024	MRIM
04/04/2014	CITIPOWER	5044050.180	77761.113	69530.046	64437.280	...	93741.476	91659.930	93286.271	96403.437	92482.282	MRIM
05/04/2014	CITIPOWER	4383318.300	80930.298	72390.242	66461.934	...	91474.783	91223.017	94194.091	98842.620	95486.413	MRIM

FIGURE 1: AEMO MRIM METER DATA FOR CITIPOWER ELECTRICAL DISTRIBUTION

The data for the Project was collected from April 2014 till September 2019, which was the latest set of data published on the AEMO website. There was data collected before April 2014. However, due to the fact that the meters were only rolled out from mid-2011, it seemed necessarily appropriate to use the data when all of the meters are properly installed i.e.. from April 2014 onwards so that the modelling is based on a consistent set of conditions without data coming from different methods of collation.

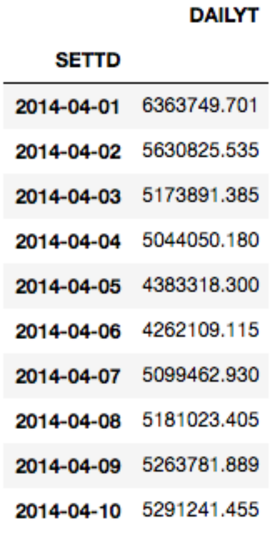
The data contained energy consumption in (kWh) which are distributed from 5 different electrical distribution networks. Each distribution network is responsible for distributing electrical energy for different parts of Victoria and hence, the behaviour of the data will most likely be different for each network. Hence, different machine learning models may perform better for each respective electrical distribution network. For this Project, the exercise will be focussed on setting an approach using the Python programming language and determining the models that will be most appropriate for forecasting the energy consumption based on Victorian MRIM meters that were installed by CitiPower.

Using the Pandas and Numpy libraries, null checks were conducted to ensure that there are no missing values in the dataset and that the data published is consistent holistically. No missing values were identified in the data. Therefore, the dataset was considered “clean”. Summation checks were also conducted to determine whether the summation of aggregated half-hour energy (VAL01 - VAL48) equals to the daily total (DAILYT). A majority of the summations were equal to the daily total. The ones that were not had an insignificant difference between the daily total and calculated summation of the half-hour energy values and this can be presumed to be just rounding errors. Therefore, it is concluded that the aggregated half-hour energy recorded each day equals to the daily total.

Additional processes on the DataFrame were conducted, as follows:

- Converted the date (SETTD) from ‘object’ to a ‘datetime’ format
- The dates were set as an index of the DataFrame.

To reduce the ‘randomness’ that can be generated from half-hourly energy values, the daily total, DAILYT, is used for the exploratory data analysis (EDA) and will be the input for the machine learning models. The end results of the data processing is shown in Figure 2.



	DAILYT
SETTD	
2014-04-01	6363749.701
2014-04-02	5630825.535
2014-04-03	5173891.385
2014-04-04	5044050.180
2014-04-05	4383318.300
2014-04-06	4262109.115
2014-04-07	5099462.930
2014-04-08	5181023.405
2014-04-09	5263781.889
2014-04-10	5291241.455

FIGURE 2

Data Analysis

The data is defined as a time series dataset, since the data is recorded in the order of time. As seen in the red graph below, a Yearly Seasonal component was observed, indicating that there is a similar pattern occurring every year. During further data decomposition, an additional weekly seasonal component was observed (see Figure 3). This indicates there are similar patterns occurring every year and week.

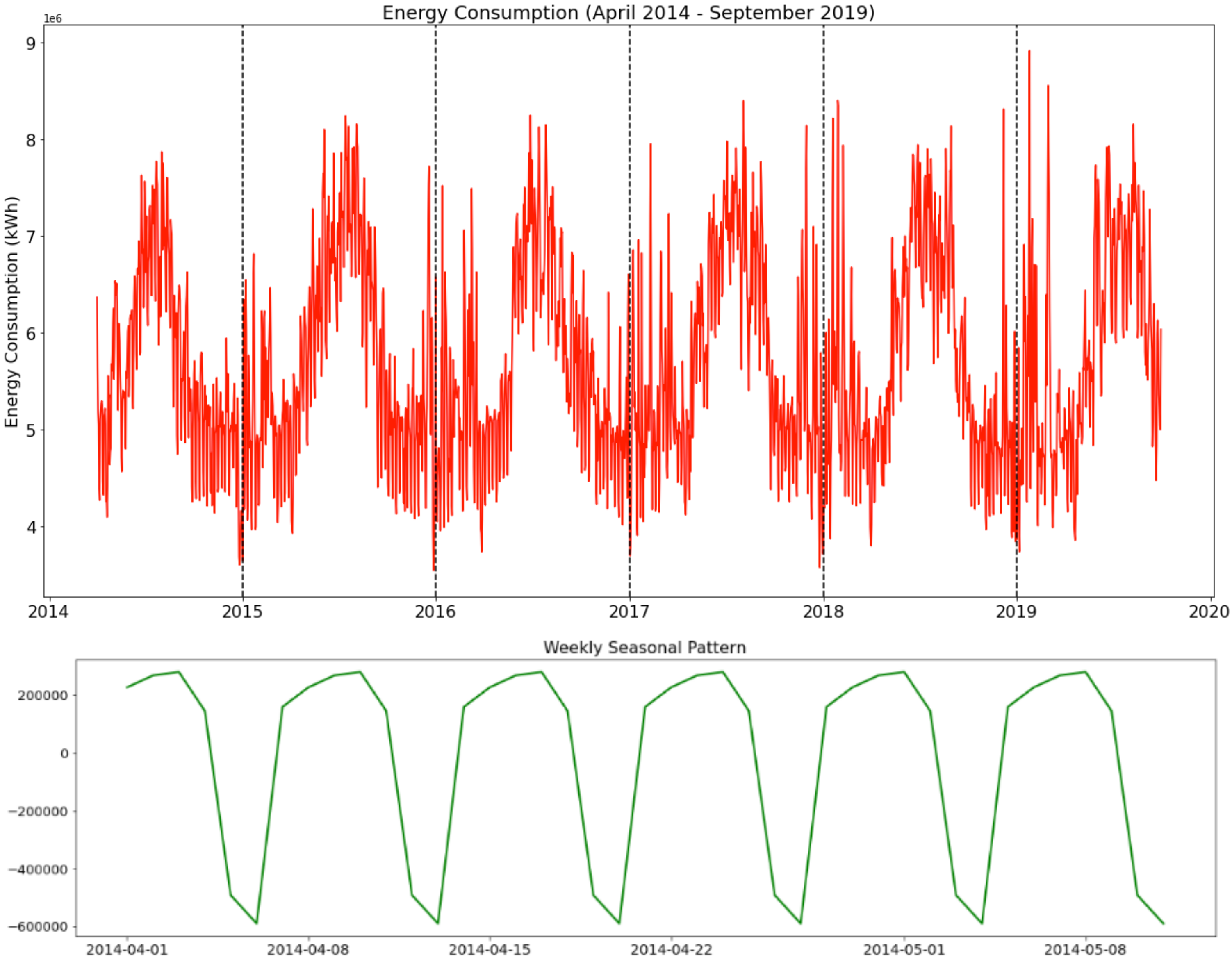
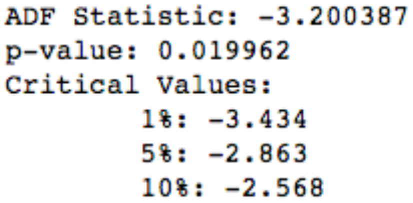


FIGURE 3: YEARLY (TOP) AND WEEKLY (BOTTOM) SEASONAL PATTERNS ARE OBSERVED FROM THE DATA

A Dickey-Fuller Test was conducted to determine the stationarity of the data. The test is a hypothesis assessment, where if the p-value from the analysis is less than 0.05, the null hypothesis is rejected and the data can be classified as stationary. The p-value for this dataset was found to be 0.019962. Therefore, the data is classified as stationary and the machine learning models utilised for forecasting will be "robust" when they are trained by using this data.



```
ADF Statistic: -3.200387
p-value: 0.019962
Critical Values:
        1%: -3.434
        5%: -2.863
        10%: -2.568
```

FIGURE 4: DICKEY - FULLER TEST RESULTS

Models

The machine learning models will be trained on the data from *1 April 2014 till 12 December 2018*. From the training data, the models will be used to forecast the energy consumption from *1 January 2019 until 30 September 2019*. The data from AEMO for this period will then be used for comparisons with the forecasts from the trained time series models.

The models that were used in the project were from the respective Python packages and are shown in the Table below:

Model	Package
ARIMA	statsmodel
SARIMA	statsmodel
SARIMAX	statsmodel
Prophet	fbprophet
LSTM	keras

TABLE 1: MODELS AND ITS SOURCE

ARIMA

ARIMA is an abbreviation for Auto Regressive Integrated Moving Average. The model captures complex relationship as it takes error forms and observations of past values. In other words, the model regresses variable on past values. It is one of the most common models used for forecasting and this provides a baseline, that is, other models should improve and provide better modelling performance.

The configuration of this model can be referred in the Jupyter Notebook: Capstone_P3_ARIMA.ipynb.

SARIMA

SARIMA is an extension of ARIMA, where the model includes an additional seasonal component for modelling. The limitations of this model is that it can only take one seasonal component. In this project, two separate models were built for a weekly and yearly seasonality respectively. As a note of caution, running the model with the inclusion of a yearly seasonality has a high computation requirement and the computer will need a high capacity to ensure that it does not crash. It is advised to run the model on a cloud computing system or a high-end computer.

The process of how SARIMA was configured can be referred in the Jupyter Notebook: Capstone_P4_SARIMA.ipynb

SARIMAX

SARIMAX is another extension of ARIMA. In addition with including a seasonal component, the model also takes in additional exogenous variables. The exogenous variable are considered external values which may assist with forecasting the energy consumption, such as the weather, or another seasonal component. For this component, an additional seasonal component was added into the exogenous variables in the form of Fourier terms. The Fourier terms was selected following research on adding seasonal components (see Reference 4). Multiple orders of Fourier terms were tried out. The 2nd order Fourier term was observed to provide a more accurate prediction.

The process of how SARIMAX was configured can be referred in the Jupyter Notebook: Capstone_P4_SARIMA_SARIMAX.ipynb.

Prophet

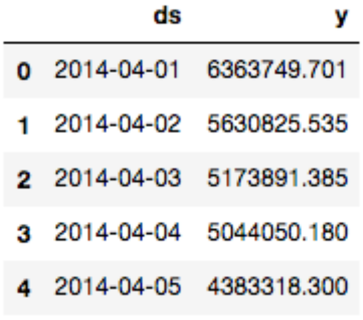
Prophet is a time series forecasting model that is developed by Facebook in 2017. It does well with time series data that have strong seasonal effects and several seasons of historical data. Furthermore, it can be fit with yearly, weekly and daily seasonality, plus holiday effects. For this model, feature engineering was done to prepare the proper input into the model to ensure that it is compatible with the machine learning algorithms requirements.

The input to Prophet is always a DataFrame with two columns:

1. ds : Date column (YYYY-MM-DD or YYYY-MM-DD format)
2. y : Measurement that is wished to be forecast

The feature engineering done towards the data were:

1. Reset the index



	ds	y
0	2014-04-01	6363749.701
1	2014-04-02	5630825.535
2	2014-04-03	5173891.385
3	2014-04-04	5044050.180
4	2014-04-05	4383318.300

FIGURE 5: INPUT FOR PROPHET

2. Renamed the column according to the requirements for the input to Prophet
    1. SETTD : ds
    2. DAILYT : y

For a better description, refer to the Jupyter Notebook: Capstone_P5_Prophet.ipynb.

LSTM (Long Short Term Memory)

Long Short Term Memory is a recurrent neural network where the model learns from a series of past observations to predict the next value in the sequence. The input is built upon where the model will forecast the next value based on 2 weeks worth of historical data. The input is then reshaped into a 3D format as expected by LSTM, namely [samples, time steps, features].

For a better description, refer to the Jupyter Notebook: Capstone_P6_LSTM.ipynb.

Additional Variation: Walk - Forward Validation

In time series modelling, the predictions over time become less and less accurate for further periods from the last dataset acquired. A more realistic approach is to retrain the model with the latest available data for further predictions. Since training of statistical models are not time consuming, walk-forward validation is the most preferred solution to get more accurate results.

This walk-forward method has thus also been implemented in the following models:

- ARIMA
- SARIMAX
- Prophet
- LSTM

All of the models and its variations are summarised in Table 2.

No	Model	Variation
1	**ARIMA**	• Basic
2	**ARIMA**	• Exact construct as Model 1 • Implemented Walk-Forward Validation
3	**SARIMA**	• Included a weekly seasonal component
4	**SARIMA**	• Included a yearly seasonal component
5	**SARIMAX**	• Used SARIMA with weekly seasonal component • Exogenous variable • Added an additional seasonal component in form of second order Fourier terms
6	**SARIMAX**	• Exact construct as Model 5 • Implemented Walk-Forward Validation
7	**Prophet**	• Basic • Added yearly and weekly seasonalities into the model
8	**Prophet**	• Exact construct as Model 7 • Implemented Walk-Forward Validation
9	**LSTM**	• Reshape DataFrame such that the model forecast the next value based on 2 weeks of historical data • Implemented Walk-Forward Validation

TABLE 2: MODELS AND ITS VARIATIONS

Model Evaluation

The performance of each model compared with the actual values, is shown in the Figure 6.

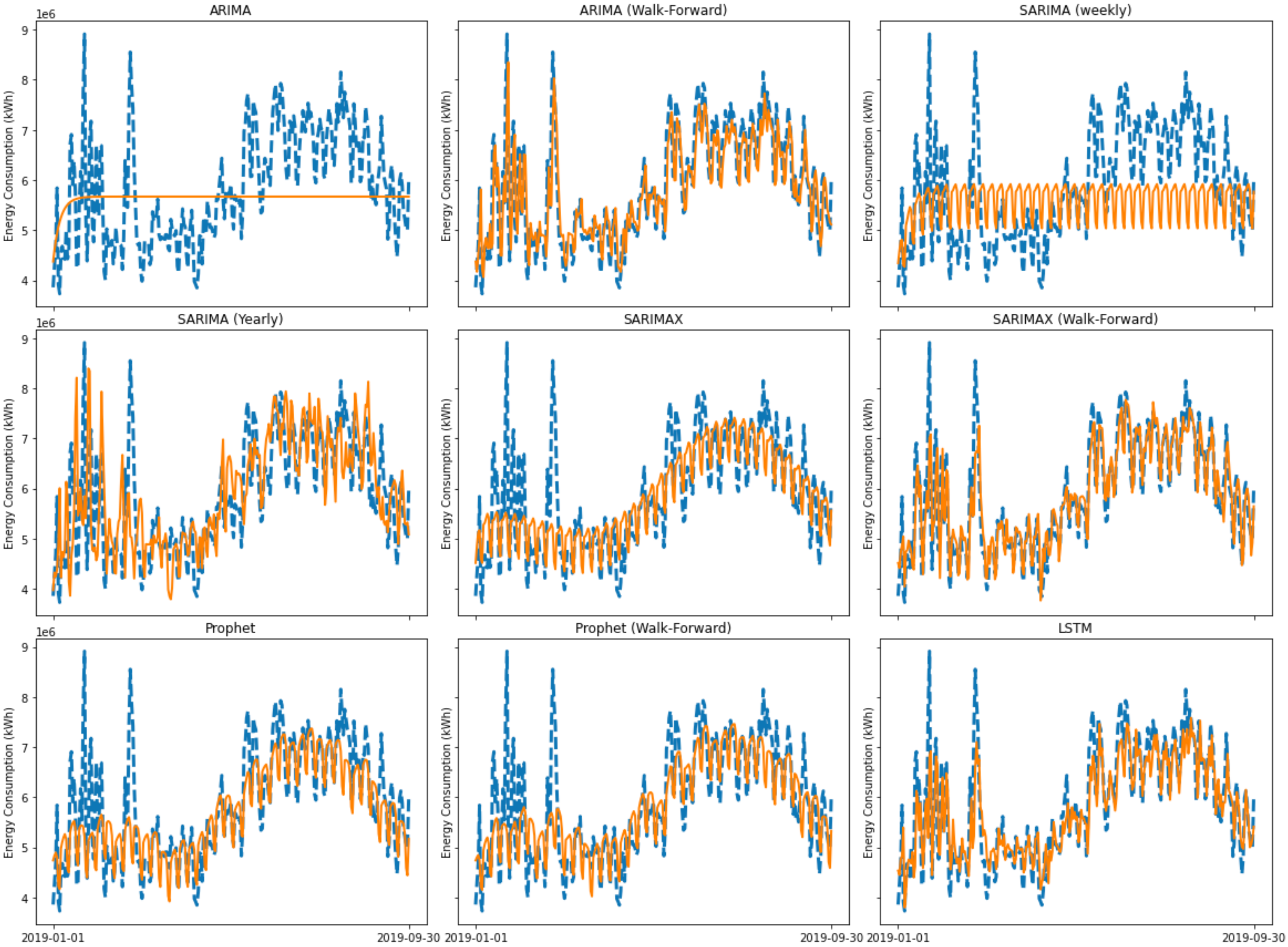


FIGURE 6: MODEL OUTPUTS (ORANGE) IN COMPARISON WITH ACTUAL VALUES (BLUE)

Based on visualisations alone, it can be seen that the top 3 performers are:

- ARIMA (Walk - Forward)
- SARIMAX (Walk - Forward)
- LSTM

The performance of each model cannot be assessed based on visualisations alone. Therefore, appropriate mathematical evaluations are used to determine the performance of these models .

The mathematical performance metrics to evaluate these models were:

- R2 Score and
- MAPE (Mean Absolute Percentage Error)

The performance metrics are compiled into a bar plot, shown in the Figure 7 and Figure 8 below.

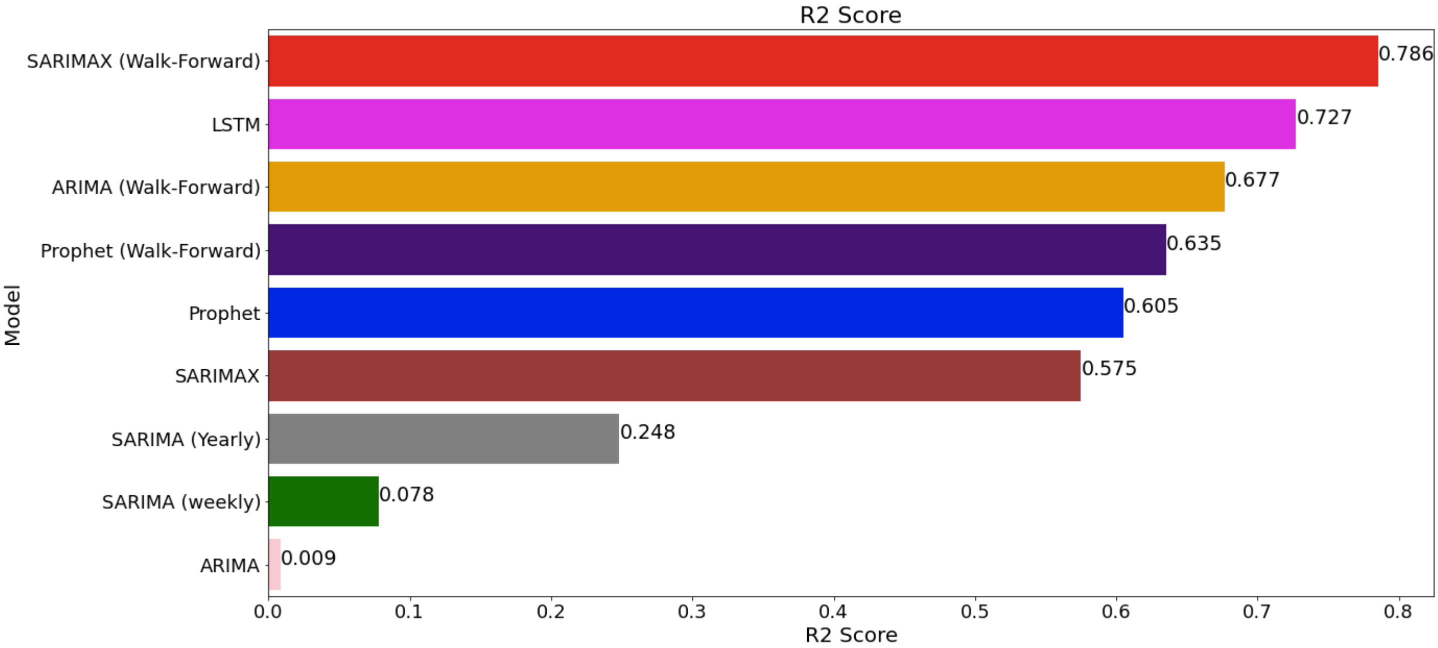


FIGURE 7: R2 SCORE

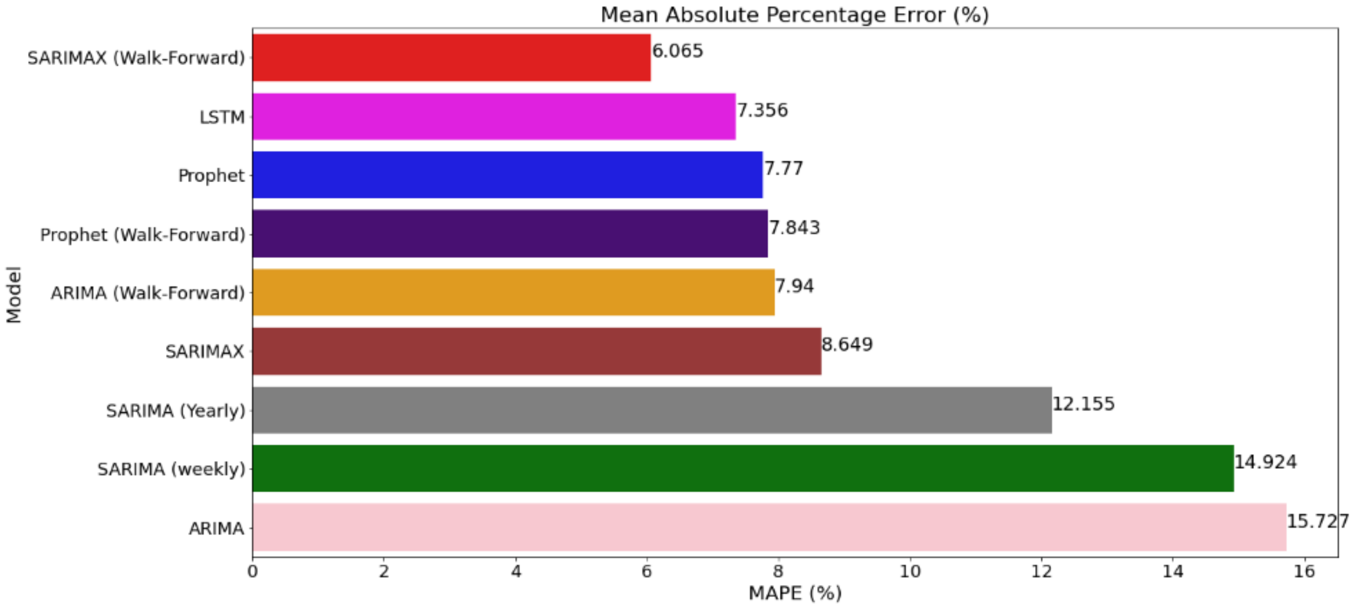


FIGURE 8: MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

Based on Figure 7 and 8, SARIMAX (Walk-Forward) was observed to perform the best in forecasting energy consumption for CitiPower, followed by LSTM. However, the comparison between the remaining models is difficult to determine. Therefore, a 2D plot was constructed where the X and Y axis represent R2 Score and MAPE respectively, to appropriately determine the models forecasting performance based upon the combination of these performance matrices The better models are represented in the bottom right quadrant of the the 2D Plot (see Figure 9).

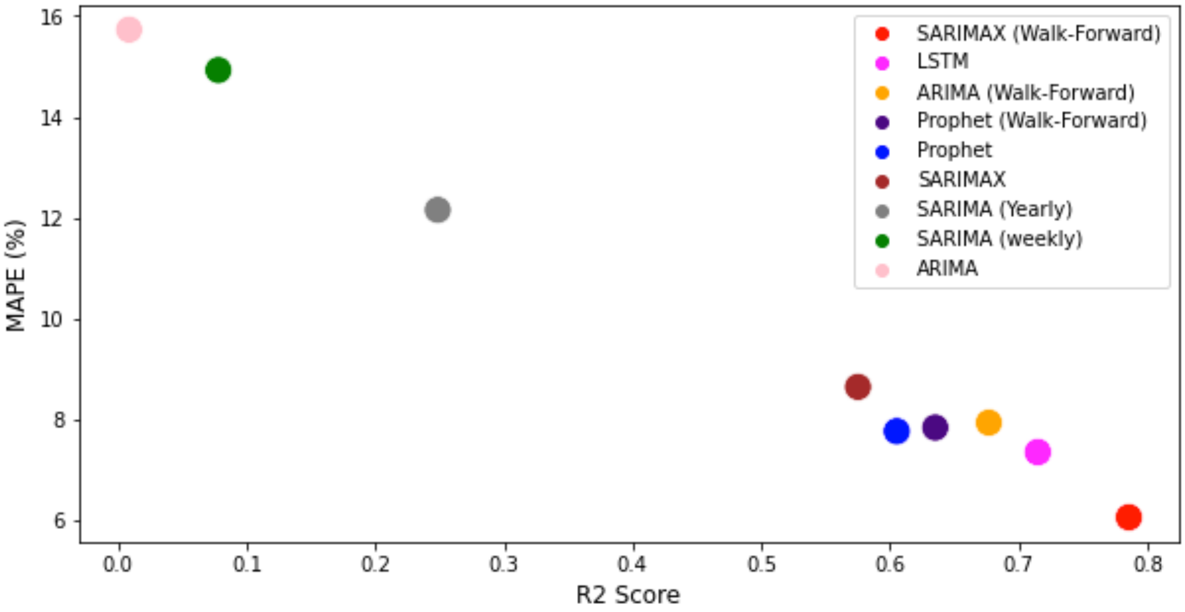


FIGURE 9

Conclusion

The Walk-Forward models are found to perform the best in forecasting the energy consumption for the CitiPower electrical distribution network. The SARIMAX (Walk-Forward) model performed the best to forecast the energy consumption from CitiPower with an R2 Score of 0.786 and MAPE of 6.065%. However, this model performs well for a short period of forecasting. For a longer period of forecasting, Prophet is the viable model with an R2 Score of 0.605 and MAPE of 7.77%.

The Project has shown that data modelling can be used for analysis and forecasting the energy consumption of an electrical distribution network. AEMO and the operator, CitiPower may be able to more effectively plan, operate and manage its electrical distribution network by applying such forecasting models for more reliable and better business outcomes.

Further Work

The SARIMAX (Walk-Forward) model was observed to perform the best to forecast the energy consumption for the electrical distribution network, CitiPower. It may be interesting to calibrate and verify models that would perform best for other electrical distribution networks.

There are additional models to be discovered within Amazon Web Services (AWS) and Microsoft Azure. It will also be interesting to discover whether these models perform better than the concluded model from the Python package in this project.

The walk-forward can be identified as a simulated system where the the model is retrained after a new data is acquired from the meters. Further work can be implemented where the an automated process is implement to retrain the model from sensors in real time.

References

1. Jupyter Notebooks
   - Capstone_P1_Compile_Datasets.ipynb
   - Capstone_P2_EDA.ipynb
   - Capstone_P3_ARIMA.ipynb
   - Capstone_P4_SARIMA_SARIMAX.ipynb
   - Capstone_P5_Prophet.ipynb
   - Capstone_P6_LSTM.ipynb
   - Capstone_P7_Evaluation_and_Conclusion.ipynb
2. AEMO I Victorian Meter Data

   https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/metering-data/victorian-mrim-meter-data
3. Jason Brownlee. “What is Time Series Forecasting?” *Machine Learning Mastery.* Aug 15 2020

   https://machinelearningmastery.com/time-series-forecasting/
4. Grzegorz Skorupa. “Forecasting Time Series with Multiple Seasonalities using TBATS in Python” *Medium.* Jan 14 2019

   https://medium.com/intive-developers/forecasting-time-series-with-multiple-seasonalities-using-tbats-in-python-398a00ac0e8a
5. Jason Brownlee. “How to Develop LSTM Models for Time Series Forecasting” *Machine Learning Mastery.* Aug 28 2020

   https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/