Comparative Study of Long-term Electricity Demand Forecasting Models

Machine Learning Final Project - Spring 2023

Geet Patel (gp507)
Computer Science Department
Rutgers University
gp507@scarletmail.rutgers.edu

Hunny Balani (hb426)
Computer Science Department
Rutgers University
hb426@scarletmail.rutgers.edu

Hasnain Gandhi (hg410)
Computer Science Department
Rutgers University
hg410@scarletmail.rutgers.edu

***Abstract*—Demand forecasting provides a problem very relevant to the current energy crisis the world faces. Accurate understanding of futuristic demand can provide a better understanding of supply chain tasks lying ahead, which helps improve the end-customer satisfaction. In this project we start by demonstrating the problem at hand, why it is important to solve and benefits it carries. We first talk about collecting data from different regions of the world. We then proceed to explain different Machine Learning and statistical models to predict the energy trend in these regions. Consequently, we provide a comparative analysis of the methods talking about the error in the prediction and how the trend augers with the actual predicted energy demand of the world. Lastly we talk about the learning's and future scope for this project.**

***Index Terms*—Forecasting, Recurrent Neural Network, ARIMA Models, regression analysis**

I. Project Overview

The project focuses on algorithms for estimating demand forecasting specifically focusing on energy prediction for different regions of the world. Demand forecasting solves important business problem for the power industry as well as for governments to plan the budgets and so has important political consequences as well. The project aims at using the time-series, regression and recurrent neural networks to predict trends.

The report is broken down into these sections. We first pledge the integrity maintained in the project code as well as the report and presentation submission. Section 3 lists down the different datasets used in the analysis which includes New York - USA, UK and Victoria - Australia. This section briefly explains the features used, features dropped, data exploratory done and the final columns kept in the project.

Section 4 goes on to explain the statistical models used in the predictive analysis. We start with briefly mentioning the ARIMA model, the importance of the parameters and how we can estimate those using different tests. We better our estimation with then explaining the SARIMA model and what effect does seasonality play in the energy estimation. We lastly explain the newer SARIMAX model which also takes in external factors like effect of rain, global shortages. Section 5 talks about the newer approaches to recurrent neural network which is LSTM and the concept is that, the feedback does help the network to learn from the previous estimations and trend. We talk about the different flavours of LSTM including the Bi-directional, vanilla and CNN-LSTM. Section 6 provides the root mean square metric to determine whether the predicted trend follows the actual trend and the deviations encountered. We can reason here which models performs better in different circumstances.

Section 7 finally talks about the learning's from the project, the future work which can better the model, the reasons of some model not performing and the reasons for that.

II. Maintaining the Integrity of the course

In the analysis of the project, we have taken reference the papers, but the code and the understanding is independent of it and the reference is used as resource for our understanding. Appropriate references are given at the end of the paper. The report and presentation write-up is based on our understanding of the project.

III. Data Set Description

For this project, we wanted to have a comparative analysis of different regions of the world and so we have used data from New York city - USA, UK and Victoria - Australia. Now we provide the description for each of the data-set.

A. New York Dataset:

The dataset consists of the following columns:

- Timestamp: Data has been recorded hourly.
- Demand: Float value denoting the energy required in each hour.
- Precipitation: Float value denoting value if rain was observed in the hour.
- Temperature: Float value in Fahrenheit.

B. *United Kingdom Dataset:*

The dataset consists of the following columns:

- Settlement Date: Data has been recorded hourly.
- ND (National Demand): Integer value denoting the energy units required in each hour.
- TSD (Transmission System Demand): Integer value which is ND plus the energy required to manage load.
- Temperature: Float value in Fahrenheit.
- Embedded wind capacity: Int value of the units generated through wind energy.
- Embedded solar energy: Int value of units generated through solar energy.

C. *Victoria - Australia Data set:*

The dataset consists of the following columns:

- Date: Data has been recorded daily.
- Demand: Float value denoting the energy units required each day.
- RRP (Recommended Retail Price): Float value mentioning the retail price for the current demand. Can see fluctuations over different days.
- Temperature: Float value in Fahrenheit.
- Embedded wind capacity: Int value of the units generated through wind energy.
- Embedded solar energy: Int value of units generated through solar energy.

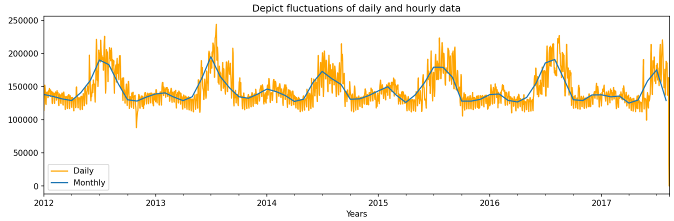
D. *Data Preprocessing*

We analysed the data and found an important observation regarding the data, that most of the entries were hourly demand entries and that would lead to potential uncertainties in the long term predictions. That motivated us to run through these preprocessing steps.

- Convert the hourly data into the daily data. This still has a lot of noise because there ought to be fluctuations in the demand per day.
- Logical next step then was to take the sum of the entire months data so that we see some trend in the demand rather than noise.
- After we have monthly demand details, we then drop columns for our prediction like the amount of rain, temperature values.
- Finally we need to separate the data into training and testing so that, we can compare the trend from our predicted algorithm to the actual data.

IV. Statistical Models for Prediction

Seasonality is basically a repetition of a particular pattern. We are studying energy demand, which also has a specific pattern around the year because of the weather or external conditions. In all the three datasets, we can observe that the demand rises during mid-year as compared to other timeframes. Thus, energy demand has some seasonal properties. We can also observe this from the plot below where we see that the trend repeats every year if we see the blue curve.



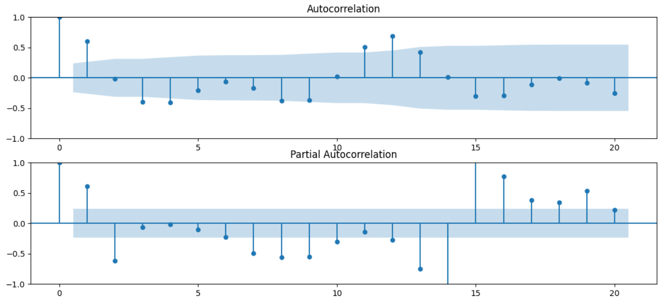
Because of the seasonal property of data, we can use statistical models such as ARIMA and its enhancements to predict future demands.

A. *ARIMA*

ARIMA can be broken down into three terms:

- AR - auto regression
- I - integrated
- MA - moving average

AR is denoted by the p parameter and MA is denoted by the q parameter, which can be determined by the ACF-PACF test. I is basically the differencing term which is a measure of stationarity, represented by the d parameter. d is basically the order of difference. The ACF plot shows the correlation between the time series and its lagged values. The PACF plot shows the partial correlation between the time series and its lagged values, which is the correlation between the time series and its lagged values. For eg, If d is 1, we will be predicting not on the original data, but on the difference of a timestamp of data.



1) *Estimating Parameters:* Now we need to find the values of p,q,d so that we can feed it into the dataset.

- Parameter d can be determined by Augmented Dicky Fuller (ADF) test.
- Parameter p is estimated by analysing the ACF-PACF plot. If the ACF plot shows a sharp drop-off after lag k and the PACF plot shows a significant spike at lag k, then the optimal value of p is probably k.
- Parameter q is estimated by analysing the ACF-PACF plot. If the PACF plot shows a sharp drop-off after lag k and the ACF plot shows a significant spike at lag k, then the optimal value of q is probably k.

B. *SARIMA*

SARIMA is basically Seasonal ARIMA. It overcomes the limitations of ARIMA by accounting the seasonal aspect of time series. Apart from the p, q, d parameters of ARIMA, we

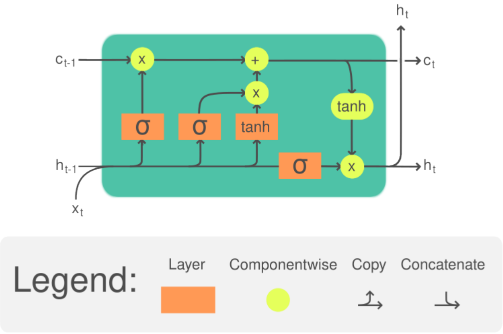


Fig. 1. Structure of an LSTM Unit

have four more paremeters that account for the seasonality. They are as follows:

- s - the number of periods in a single seasonal cycle
- P - number of autoregressive terms in the seasonal component of the model
- Q - number of moving average terms in the seasonal component of the model
- D - number of seasonal differences needed to make the time series stationary

C. SARIMAX

SARIMAX is basically Seasonal ARIMA + eXogenous factors. In this case, the exogenous factors are temperature, price, precipitation, etc. We have used auto_arima function provided by pmdarima library. This function helps in selecting all the hyper parameters and the perfect model best suited for training.

This function used information criterion values to select the best parameters. We have used AIC, hence the parameters for which the AIC is lowest is taken into consideration. With auto aroma, the best model we could find was SARIMAX.

V. LSTM Approach

LSTM are a special type of neural networks. They are primarily an extension to recurrent neural networks (RNNs) which, unlike feedforward networks, also contain feedback connections. This property translates into improvements in sequence modelling tasks such as natural language, speech processing and data prediction. But over RNNs, each unit of LSTM consists of trainable input, output and forget gates. This unit structure can remember values over arbitrary time intervals and the three gates regulate ho the information is flown in and out of the cells. Each gate is associated with a value between 0 and 1, which controls the flow of information from that gate.

A diagram is shown in figure 1. Moving from left to right, first σ and $\times$ form the forget gate, further right the second $\sigma, \tanh, \times$ and $+$ are the input gate, and the rest set up the output gate.

A. Different structures

We can use this unit to build a simple structure with one LSTM and one linear layer, which is known as vanilla LSTM. We can also stack multiple of these units together to create a Stacked LSTM. An eclectic hierarchy can be created as CNN-LSTM, in which a convolutional layer is added before the LSTM. Derived from RNNs, one more approach is called Bi-LSTM, in which information is flown from not only $t-1$ to t, but also from $t+1$ to t as well.

VI. Results and Predictive Analysis

So now, we present the RMSE values for the different statistical models, and try to reason the best approaches.

	ARIMA	SARIMA	SARIMAX
NYC	0.0607825	0.0557089	0.0311674
UK	0.0725248	0.0358596	0.0181307
Victoria	0.0280235	0.0141515	0.0177171

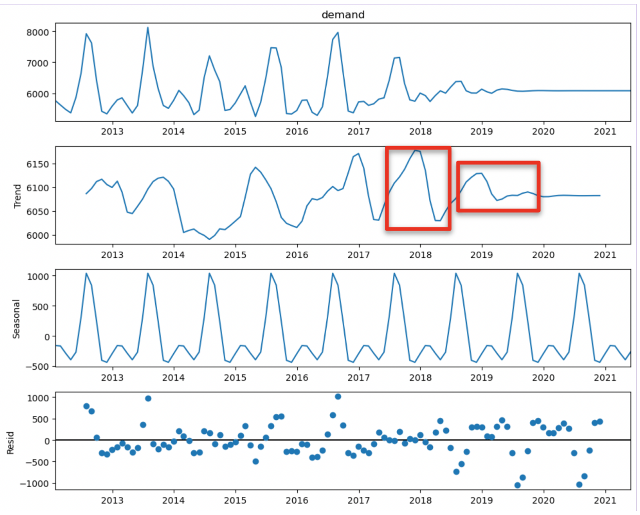
TABLE I
Root Mean squared Error for Different Methods

From the table we can gauge that the rmse values progressively decreases if we increase parameter, in other words, SARIMAX with seasonal as well as exogenous factors provides a good estimation of the prediction. Another factor which could be attributed to it would be that the autoarima function inherently estimates the parameters so the accuracy to find the correct parameters does increase. Now let us try to reason what the graphs of prediction tell us about the nature of the dataset and map it to the actual world demand during pre and post COVID times.

For that we predicted the energy consumption for all the datasets up to certain years and drew conclusions from actual energy consummations and future projections. The predictions were as follows:

A. New York

For NY, we had the data from 2012 to 2017. We predicted the energy consumption until 2021. Our predictions did support and observe a similar trend which were very similar to the actual data.

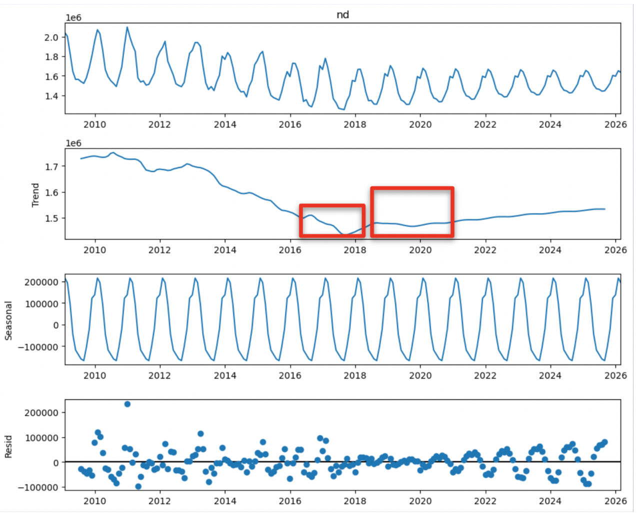


We had the following observations:

- We did see a decline in our prediction which was similar to the case with covid.
- There was a global reduction in energy demand across the globe in 2018, which also hold true for the prediction.

B. United Kingdom

For UK, we had the data from 2009 to 2023. We predicted the energy consumption for the years 2018 to 2026. Our predictions did show a similar trend, we saw a decline for 2018 and 2020.

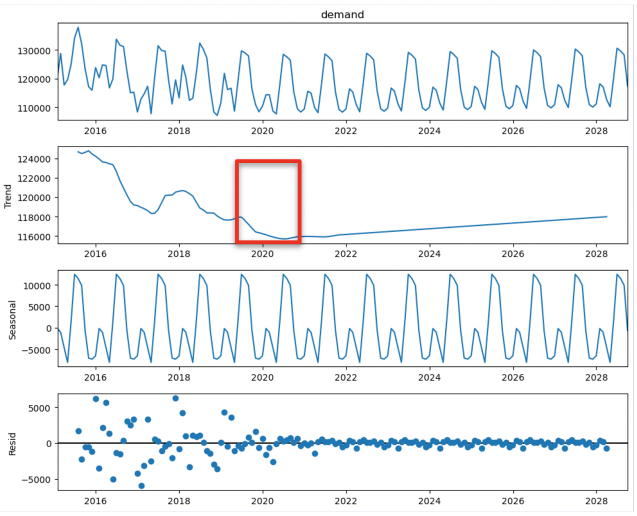


We had the following observations:

- We saw the decline during the covid period which aligns with our prediction.
- The UK demand did go down significantly as seen in many articles in 2018.

C. Victoria, Australia

For Victoria, we had the data from 2015 to 2020. We predicted the energy consumption for the years 2021 to 2028. We found the energy consumption projections for Victoria set here in Table C.1.



We found the projection of demand for Victoria state up to 2030. Our predictions do exhibit a similar pattern for the predicted years.

VII. Learnings and Future Work.

We did see that statistical models such as ARIMA can perform as well as machine learning models. This was interesting to see that statistical models can also be used for such time series analysis.

We tried to implement many different kinds of models of LSTM, however, the sky is the limit. We thought of some new approaches such as Multi-Seasonal Net - LSTM that uses various decompositions of the same dataset. The problem provides such real world consequences that it has intrigued research for many methods and we believe that different RNN based model would lead the way in those forecasting and perfectly map the trend.

References

[1] Ibrahim El-Amin Tawfiq Al-Saba. In *Artificial neural networks as applied to long-term demand forecasting*, 1999.

[2] Mostafa Shabani Hossein Abbasimehr. An optimized model using lstm network for demand forecasting. 2020.

[3] Brendan Artley. Time series forecasting with arima , sarima and sarimax. 2022.

[4] Timothy Goodson Stephanie Bouckart. The mysterious case of disappearing electricity demand. 2019.

[5] Reuters Staff. U.s. electricity use to drop by record amount in 2020 due coronavirus. 2020.

[6] Simon Evans. Analysis: Uk electricity generation in 2018 falls to lowest level since 1994. 2019.

[7] Iain Staffell Daniel Mehlig, Helen ApSimon. The impact of the uk's covid-19 lockdowns on energy demand and emissions. 2021.

[8] Iain Staffell Daniel Mehlig, Helen ApSimon. Covid-19 impact on electricity.

[9] Matthew Foley Sebastian Drimer Glenn Currie, Alexander Diplaris. Dataset for the victorian energy transition including technical, social, economic, and environmental detail. 2023.