ELECTRICITY LOAD FORECASTING AND ANALYSIS USING MACHINE LEARNING TECHNIQUES

A PROJECT REPORT

Submitted by

BL.EN.U4EEE16068 SARADHI S

in partial fulfillment for the award of the degree 0f

BACHELOR OF TECHNOLOGY

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June-2020

AMRITA VISHWA VIDYAPEETHAM
AMRITA SCHOOL OF ENGINEERING, BANGALORE, 560035

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **ELECTRICITY LOAD FORECASTING AND ANALYSIS USING MACHINE LEARNING TECHNIQUES** submitted by

BL.EN.U4EEE16068 SARADHI S

in partial fulfilment of the requirements for the award of the Degree **Bachelor of Technology** in **Electrical and Electronics Engineering** is a bonafide record of the work carried out under my(our) guidance and supervision at Amrita School of Engineering, Bangalore.

Ms.Syama S
Assistant professor
Dept. of EEE
Amrita School of Engineering
Bangalore.

Dr Rashmi M R
Chairperson
Dept. of EEE
Amrita School of Engineering
Bangalore.

This project report was evaluated by us on .....

EXAMINER 1 EXAMINER 2

Date: 16-06-2020

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Saradhi Somarouthu, S/O- Venkateswara Rao, a student of Electrical and Electronics Engineering from Amrita Vishwa Vidyapeetham, Bangalore has completed internship for a client project which involved the skill of Forecasting, which is an on-going solution development activity within the company. His project manager was Mr. Harsha Varun and he is working on this project from 13 January 2020. His employee ID here during internship was I0136. In this period his services were found to be satisfactory.

We wish him all the best in his future endeavours.

For BRIDGEi2i Analytics Solutions Pvt Ltd.

Preeth Joseph,
Director – Talent Management

ACKNOWLEDGEMENT

I offer our sincere pranam at the lotus feet of Universal guru, **MATA AMRITANANDAMAYI DEVI** who showered her blessings throughout this project.

I am very thankful to **Br.Vishwamrita Chaitanya Swamiji , Director**, Amrita School of Engineering, Bangalore, for his valuable support.

I am indebted to thank **Dr Nagaraja S.R., Associate Dean**, Amrita Vishwa Vidyapeetham, Bangalore for engraving a path for us to utilize the available resources to the fullest and thereby widen our perspective of education and growth through it.

I deeply express our sincere thanks to **Dr Rashmi, Professor and Chairman,** and **Dr. Mini Sujith, Vice-Chairperson,** Department of Electrical and Electronics Engineering for giving us the opportunity and encouraging us to use our abilities to the fullest in doing the project.

It is our privilege to express our sincere regards to our project guide, **Ms. Syama S, Assistant professor**, Department of Electrical and Electronics Engineering, for the valuable inputs, guidance, encouragement, wholehearted cooperation, and constructive criticism throughout our project.

I wish to thank our lab tutors and non-teaching staff for their undivided support with which I couldn't have completed our project.

ABSTRACT

Forecasting of electric power demand is considered as essential prerequisite in the electricity generation, transmission and distribution industries. It enables the foundation and influences the decision making process in power systems design, planning and operation. Numerous methods for forecasting and predicting electricity consumption demand are in place and operational by both public held and private energy production and distribution companies, which are very broadly classified as short-term, medium-term or long-term forecasting.

It is predominantly evident that the variation of electricity consumption demand arises due to complex interactions between meteorological and socio-economic factors such as temperature, population, cost of electric power. In such a varied dynamic environment factors ordinary forecasting and prediction algorithms, techniques have failed to produce efficient and accurate results. Hence the need for more sophisticated techniques arises.

The objective of this project is to efficiently use historic data over past 15 years to observe trends, seasonality and leverage the extensive historic data to forecast and predict electricity consumption demand. The time series analysis ARIMA and a robust machine learning algorithm- Extreme Gradient Boosting (XGBoost) has been used to perform the forecasts. The mentioned two algorithms are modelled separately and the results are compared and analysed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND NOMENCLATURE

Abbreviation	Full form	Page no.
CSV	comma-separated values	3
ARIMA	Auto Regressive Integrated Moving Average	5
XGBoost	eXtreme Gradient Boosting	5
FIS	fuzzy inference system	5
ML	Machine Learning	5
MLE	maximum likelihood estimation	5
LLE	locally linear embedding	5
EDA	exploratory data analysis	8
ADF	Augmented Dickey Fuller test	24
ACF	auto-correlation function	24
PACF	partial autocorrelation function	24
MAPE	mean absolute percentage error	30
MW	Mega Watt.	31
AR	Auto Regressive	23
MA	Moving Average	23
ARMA	Auto Regressive Moving Average	23
SARIMA	Seasonal Autoregressive Integrated Moving Average	38
AIC	Akaike information criterion	35
BIC	Bayes Information criterion	35
CSS	conditional sum of squares	35
SCADA	supervisory control and data acquisition	38

CHAPTER 1

INTRODUCTION TO ELECTRIC POWER LOAD FORECASTING

1.1 Background

During the mid-1980's the electricity energy sectors were highly regulated in majority of the countries across Asia, Europe and United States of America. Utility monopolies and distribution companies used short-term load forecasts to make sure the reliability of power system supply and long-term load demand forecasts as the foundation for planning and investing in new capacity power production plants. However, since the early 1990s, the process of de-regulation and the growth of competitive energy sector markets, integration of renewable energy, smart grids have disrupted the landscape of the traditional government owned power sectors. At the corporate level, electricity load and price forecasts have become important prerequisite for production and distribution companies' decision making processes. The loss of overestimating or underpricing and buying electric power in the real-time are possibly large, that it may lead to huge financial losses. In this scenario electric utilities, distribution companies are the most vulnerable, since they cannot levy their losses on to the retail customers. Hence optimization of losses by accurately forecasting electric load power demand have grown to be utterly essential. Today Forecasting electricity power demand is quite matured and sophisticated. Short-term forecasts that include prediction to few minutes, hours, or days ahead to the long term- up to 20 years ahead forecasts have become growingly important since the deregulation of centralized power systems. In 2000's Countries have followed isolationism and privatized a part of their power sectors, and electricity has been turned into an important commodity that can be forecasted and priced accordingly available at market.

1.2 Factors influencing Load Forecasting

It remains very essential to have a deeper understanding of the factors that determine the electric power consumed in a region over a period of time. The following are the most influencing factors and a brief overview on their affect.

1.2.1 Meteorological factors

Meteorological factors include temperature, weather conditions such as snow, rain, hail storm, global average temperature of earth, regional average temperature, latitude and longitude. Factors such as average day sunlight, cloud trends, wind speed and humidity in the atmosphere also influence variation in the consumption of the electric power load. Table [1.1] indicates the correlation of temperature and load.

Table. 1.1 Correlation of temperature and Load [1]

		Daily electricity load (L)	Maximum Temperature (MaxT)	Minimum temperature (MinT)	Sunny/rainy (SR)
Pearson correlation	L	1.000	-.237	-.588	-.527
	MaxT	-.237	1.000	.683	-.052
	MinT	-.588	.683	1.000	.581
	SR	-.527	-.052	.581	1.000
Sig.	L		.113	.000	.002
	MaxT	.113		.000	.397
	MinT	.000	.000		.001
	SR	.002	.397	.001	
N	L	28	28	28	28
	MaxT	28	28	28	28
	MinT	28	28	28	28
	SR	28	28	28	28

1.2.2 Time of the Day

The electric power consumption is not same at various time periods of a 24 hours' time span. That include time of the day, weekday/weekend, holidays, special days/festivals and seasons. It has been observed that there are significant differences between the peak and off peak load demand during a particular time span in the

traditional grid load profiles. This parameter is utmost crucial in short term forecasting. Fig. 1.1 depicts the variation in the cumulative hourly consumption in a day from data which is in csv format.

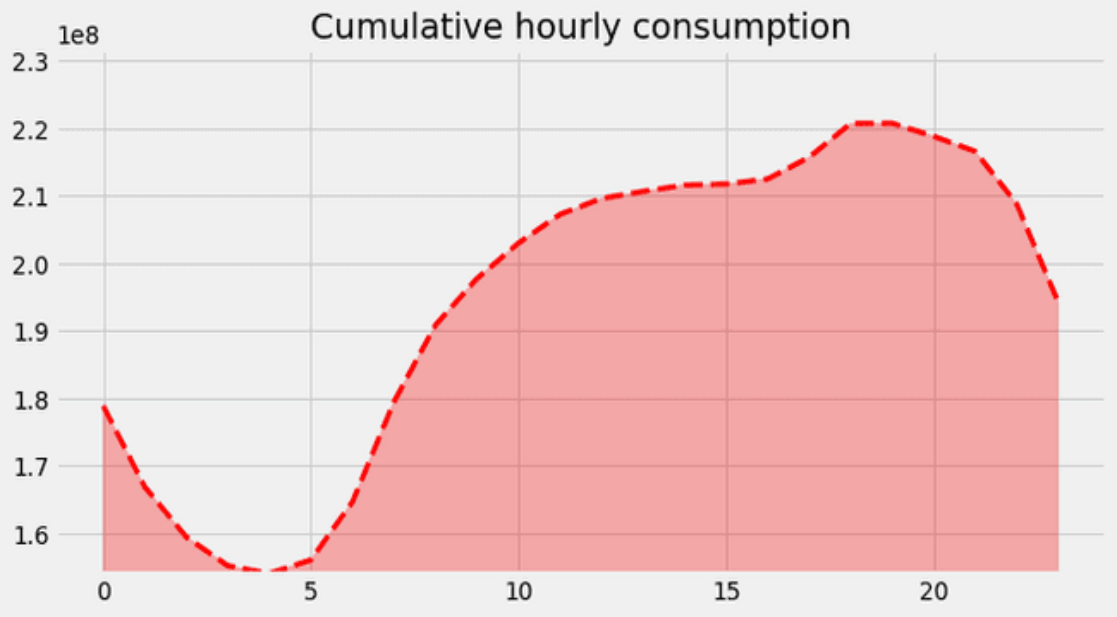

Fig. 1.1 Cumulative hourly consumption

1.2.3 Electricity prices

Distribution companies provide flat tariff rates to consumers for their daily consumption in a regulated market. Now it is possible with the integration of smart digital meters, for the consumers to respond to distribution companies' tariffs. Utilities, therefore are implementing various tariff slabs such as TOU, Critical Peak Pricing for the electric power demand. The TOU pricing provides least off peak rates but high peak cost, which will force the consumers to shift the peak loads to the off-peak loads. The amount of load that the customer shifts to the off peak times depends on the price that utility is offering at that time interval.

Ultimately effecting of unit price on the demand can be proved as a defining parameter for spot electricity price and demand mutually affect each other in the smart grid

environments that are integrated with renewable power such as wind energy and solar energy.

1.2.4 Other factors

The power load demand depends on various other factors that are listed below:

- Demand Response
- Renewable Energy Sources
- Storage Cells
- Random Disturbances
- Electric Vehicles
- Smart Grid implantation

1.3 Forecasting horizons / Methodologies of Forecasting

1.3.1 Medium/Long-term Load Forecasting Methods

End use and Econometrics are two methods used for medium and long-term forecasting analysis. Approach The end use method directly forecasts energy consumption by using extensive information on the final consumer, such as machines, the customer use, their age, sizes of houses, and so on. End use algorithm looks on the multiple uses of electricity in the domestic, commercial use, and industrial sector. These methods are derived from the concept that electric load demand is result from consumer's demand for lighting, cooling, charging vehicles, heating, refrigeration, etc. The econometric method is a combination of statistics with economic theory forecasting electricity demand. The method predicts the relevance between electric load consumed i.e. dependent variable and factors affecting the load consumption.

1.3.2 Short–term Forecasting Methods

Short-term forecasting is very essential for distribution companies and utilities to maintain stability and reliability in power systems. Short-term forecasts include prediction to few minutes, hours, or days ahead. Various methods of forecasting are depicted in fig. 1.2. The various methods used in short-terms are listed as follows:

- Similar-day Approach
- Regression Methods
- Time Series
- Expert Systems
- Fuzzy Logic
- Neural Networks

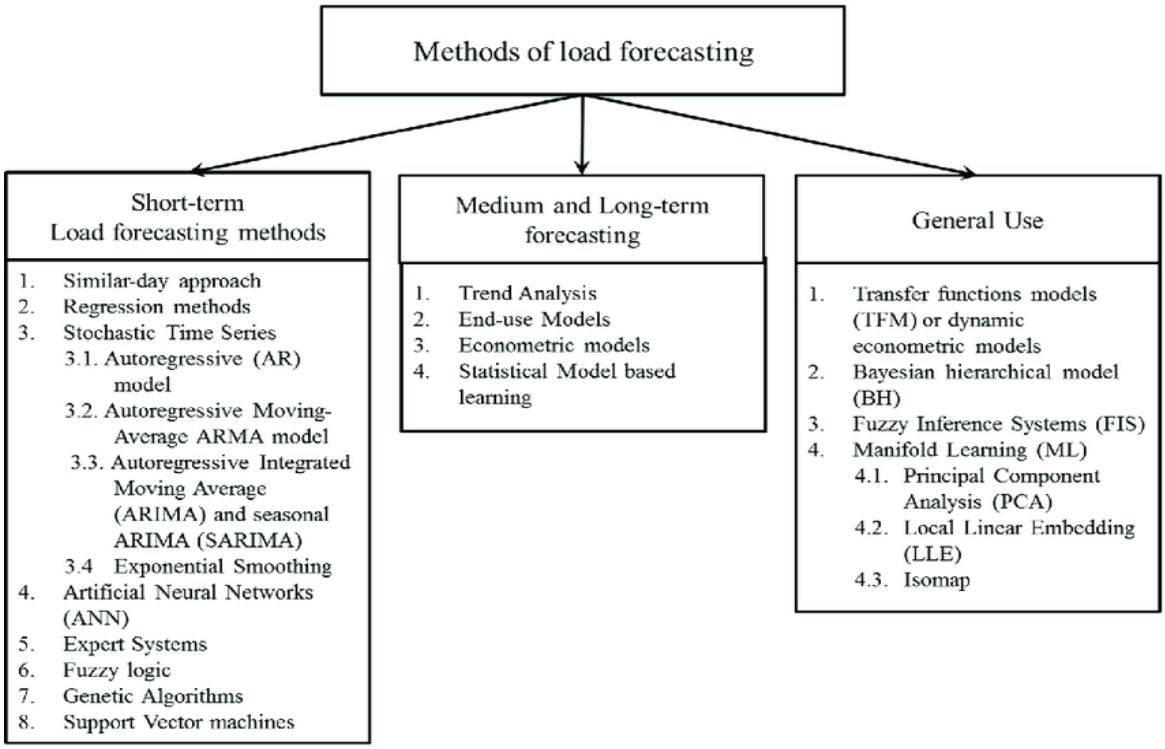

Fig. 1.2 Methods of load forecast [2]

1.4 Advantages and Disadvantages of Various methods

The advantages and disadvantages of the various forecasting methods are tabulated in table 1.2.

Table. 1.2 Comparison of various forecasting method [3]

Methods	Advantage	Disadvantage
ARIMA/State Space	• General class of nonlinear model used for forecasting a regression model and developing a fit. • Parametric and autoregressive model used for forecasting applications	• Need Data linearization. • Can only work with stationary data. • Required complex data pre-processing. • Loss of information in residuals for periodic data determined by autocorrelation and partial autocorrelation graphs. • Need complex differencing, deflating, and logging techniques for data linearization.
Decision Tree	• Require less effort for data preparation. • Tree performance is not effected by nonlinear data. • Easy interpretation	• Required more memory. • Difficult to prune. • High computational time.
FitNet	• Simplicity, improved forecast accuracy. • Autocorrelation remains constant over time.	• Nonrecurring, constant data set, non-exogenous feedback, • solution rely on non-important parameters and less. • Might over fit. • High MAPE values.
NARX-LSA-EWD	• Dynamic recurrent network and create memory between inputs and output through lags. • Faster convergence. • The accuracy is improved by utilizing optimized LSA	• More computational time

1.5 Conclusion

In this chapter, the history and background of the power load demand forecasting have been discussed and its evolution from non-existing in a regulated sector to sophisticatedly developed mythologies in an unregulated sector is reviewed. Later part of the chapter elaborated the factory or the dependent parameters that affect the variation in electric power load forecasting was discussed. It is essential to discuss these parameters since and deeper understanding these dependent parameters will help us forecast accurately and efficiently. At the end various methods that are currently used are discussed. Over the next few chapter following two methods to forecast will be discussed.

- ARIMA
- Extreme gradient descent Boosting (XGBoost)

The results of these two algorithms will be compared and analyzed in the final chapters.

CHAPTER 2

DATA INTERPRETATION

2.1 DATA OVERVIEW

Any machine learning and time series forecasting problem essentially begins from understanding the data available to us. The first and foremost step in analysis is to draw valuable insights from data like understanding the trends, independent variables, exogenous parameters and the logic in defining the dependent variables. The data is hourly electric load consumption data from a region that has been aggregated for a span of each hour. The average load for each hour that is 24 samples for each day is available from January 1 2002 to December 31 2014.

The data is very extensive with 140160 rows. This data is more than sufficient to provide us historic insights into the load consumed and will enable us to make accurate and effective forecasts. It is constituted of two columns the index being the time stamp and the second column being the load consumed in Mega Watt hours.

Fig. 2.1 is a small snapshot from the data set.

Date time	Load_in_Mw
31-12-2002 01:00	26498
31-12-2002 02:00	25147
31-12-2002 03:00	24574
31-12-2002 04:00	24393
31-12-2002 05:00	24860
31-12-2002 06:00	26222

Fig. 2.1 Snapshot from data set

The data will be split into train and test data for validation in the machine learning algorithm and the test data will be used to validated the prediction made by the proposed model. In the time series forecast will perform multilevel out of the sample forecasts assuming that the data doesn't exist. This will be discussed in chapter 4.

The salient features on the data are

- Extensive data over large time period.
- The sample time is one hour which is quite significant for 16 years of data.
- Low missing values.
- Sufficient data to perform extensive validations.

Fig. 2.2 is the graph of the load consumed with respect to the time over the years 2002 to 2012. The x-axis is time and y-axis is he load consumed in Mwh.

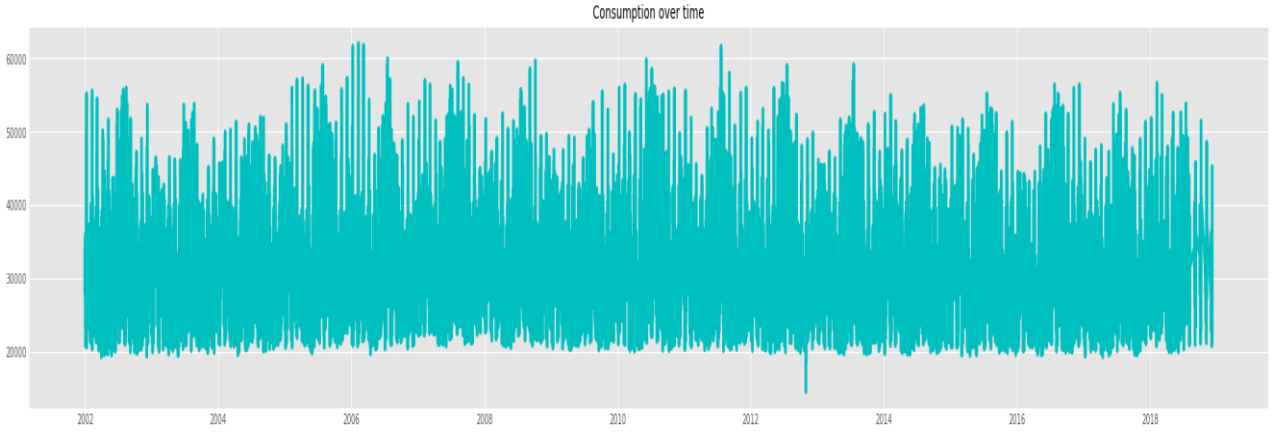

Fig. 2.2 Consumption over time

2.2 EXPLOLATORY DATA ANALYSIS

Exploratory data analysis is the preliminary step of analyzing the data set to determine the main salient features, spot anomalies, figure out patterns, check assumptions and also the secondary features that are often associated with the data. These feature are not directly determinable and often requires business context of the problem statement in hand. Exploratory data analysis extensively makes use of visual tools.

The primary objectives of Exploratory data analysis are:

- Provide hypotheses regarding the cause of the phenomenon.

- Make reasonable assumptions bases on data.
- Provide the foundation for further processes.

The tools and techniques used for EDA are:

- Box plot
- Histogram
- Scatter plot
- Line chart
- Heat map

Fig. 2.3 is the histogram of all the samples from the data. It is observed that the distribution plot is slightly left-skewed, with the most of the consumption in the range of 25,000 to 35,000

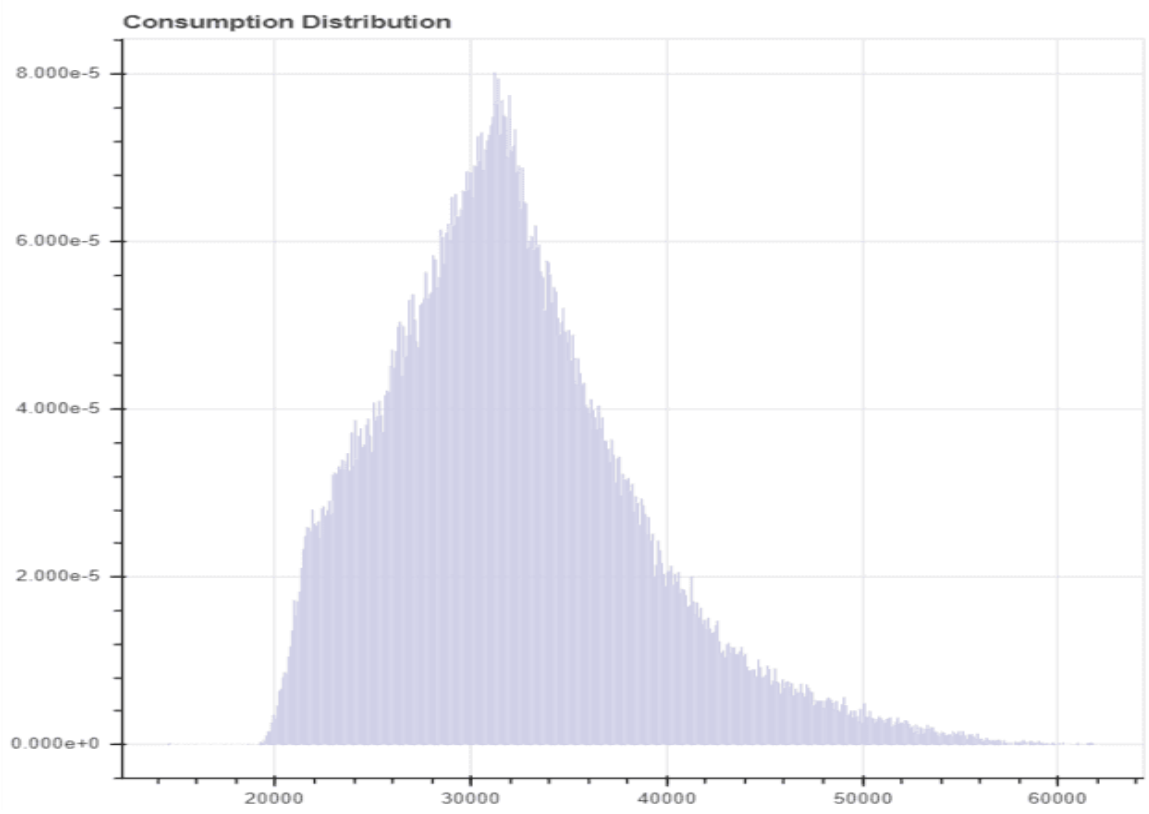

Fig. 2.3 distribution Histogram

Fig. 2.4 depicts the consumption distribution over the quarters of the calendar year. Quarterly analysis is important for companies to assess the decisions because the financial decisions are based on quarters.

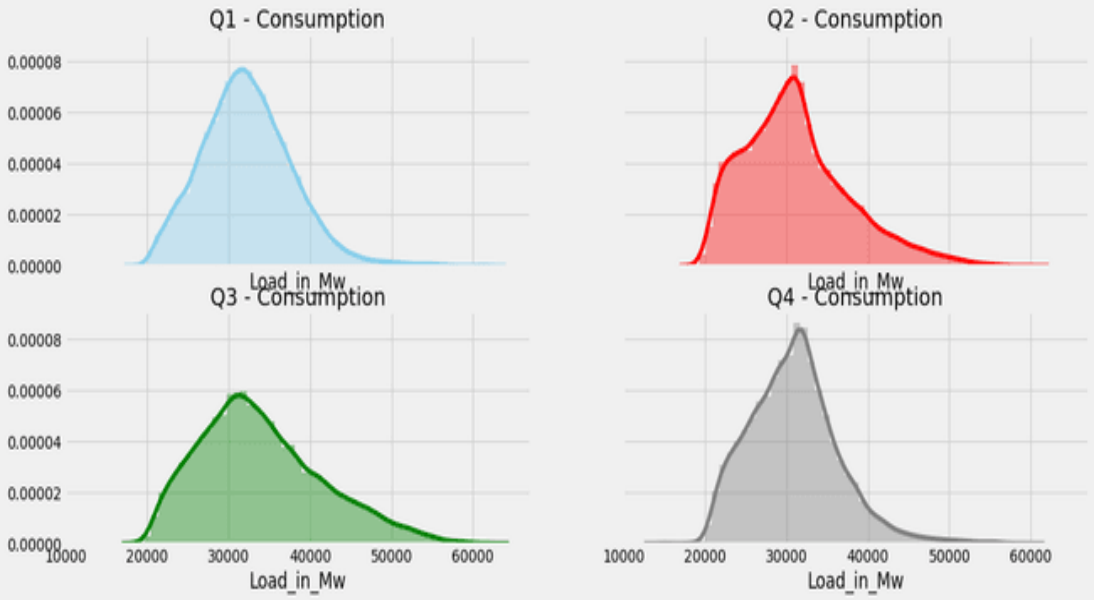

Fig. 2.4 Quarterly distribution Histogram

Fig. 2.5 plot describe the overall trend in the historic data over 16 years. It can be clearly observed that there is an upward trend. Fig. 2.6 is the scatter plot of the same trend.

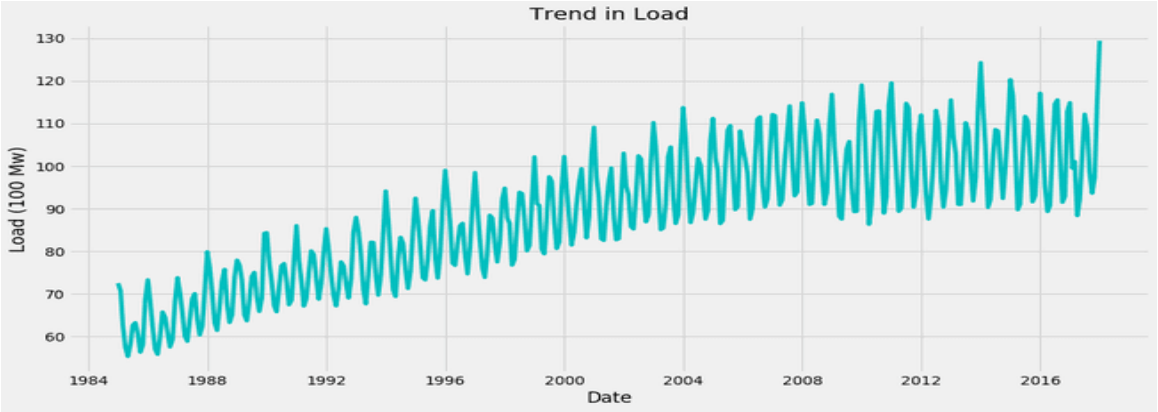

Fig. 2.5 Trend plot

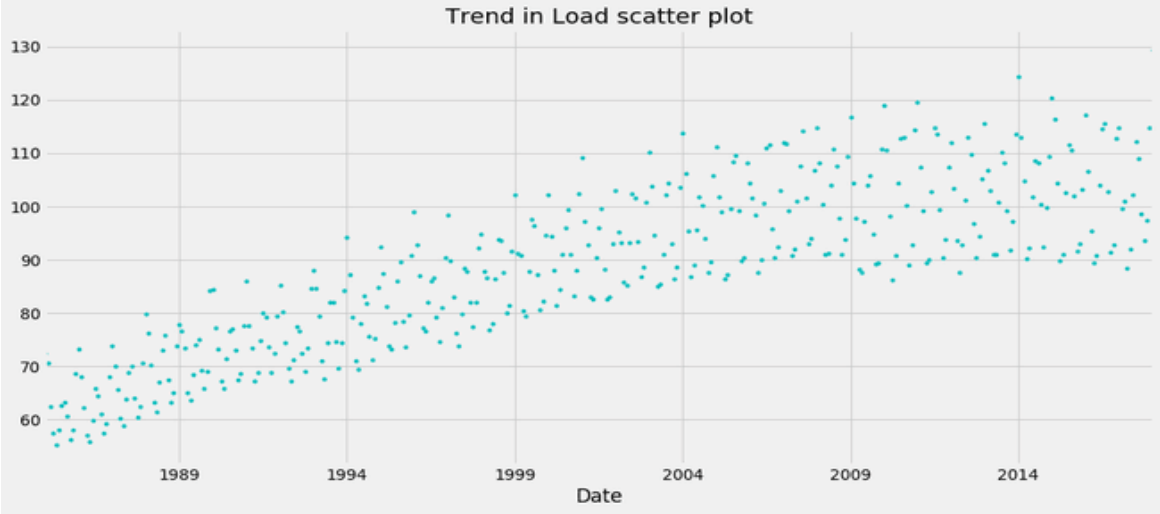

Fig. 2.6 Scatter plot

2.3 POWER SYSTEM QUALITIES

To understand the power system qualities of the data, two samples of data of year 2006 and 2015 and calculated the power system factors such as demand factor, load factor, average demand, maximum and minimum demand for both the years. Fig. 2.7 is the load duration curve of 2007.

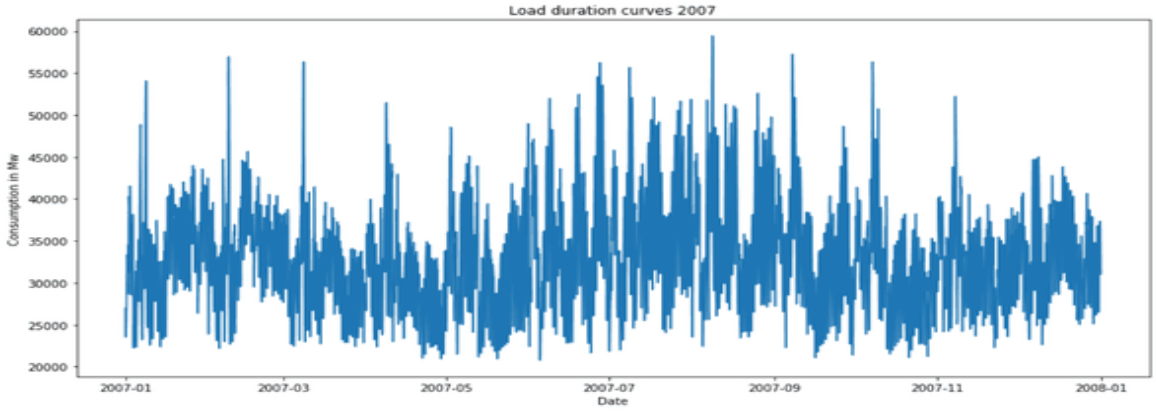

Fig. 2.7 Load duration curves 2007

The yearly average, maximum demand, load factor, base load and demand factor as Follows.

Yearly average: 31717.030267888207

maximum demand: 55129

Load Factor: 0.5753238815847957

Base Load: 19450

Demand factor: 0.7754323658509135

This analysis for the year 2017.The load duration curve is plotted and the above mentioned power quality factors have been calculated.

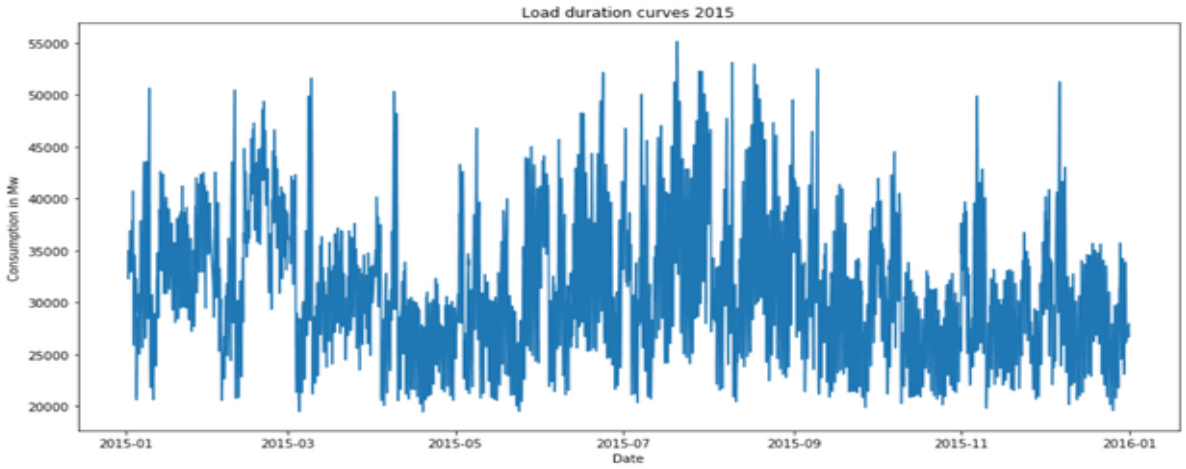

Fig. 2.8 Load duration curves 2015

These Yearly average: 33614.36428000457

maximum demand: 59437

Load Factor: 0.5655461123543344

These parameters such as demand factor, load factor and the load duration curves enables us to get a better overview of the business context of the data. Fig. 2.8 illustrates the load duration curve for 2015. The values of the demand, load factors approach one in ideal case. Load factor of 0.56 implies that 56% of the maximum load is being consumed.

CHAPTER 3

MACHINE LEARNING APPROACH- XGBOOST

3.1 INTRODUCTION TO XGBOOST

XGBoost is a sophisticated ensemble machine learning algorithm that uses decision trees based approach and gradient boost framework. In problem statements that constitutes of data that is not well structures such as images, audio, text, signals techniques such as neural networks tend to perform tremendous effective.[11] But in problems with well-structured data such as numbers and CSV files decision tree based techniques are the industry standards. The fig. 3.1 clearly elaborates the evolution in the decision trees, random forests and gradient boosting, bagging techniques.

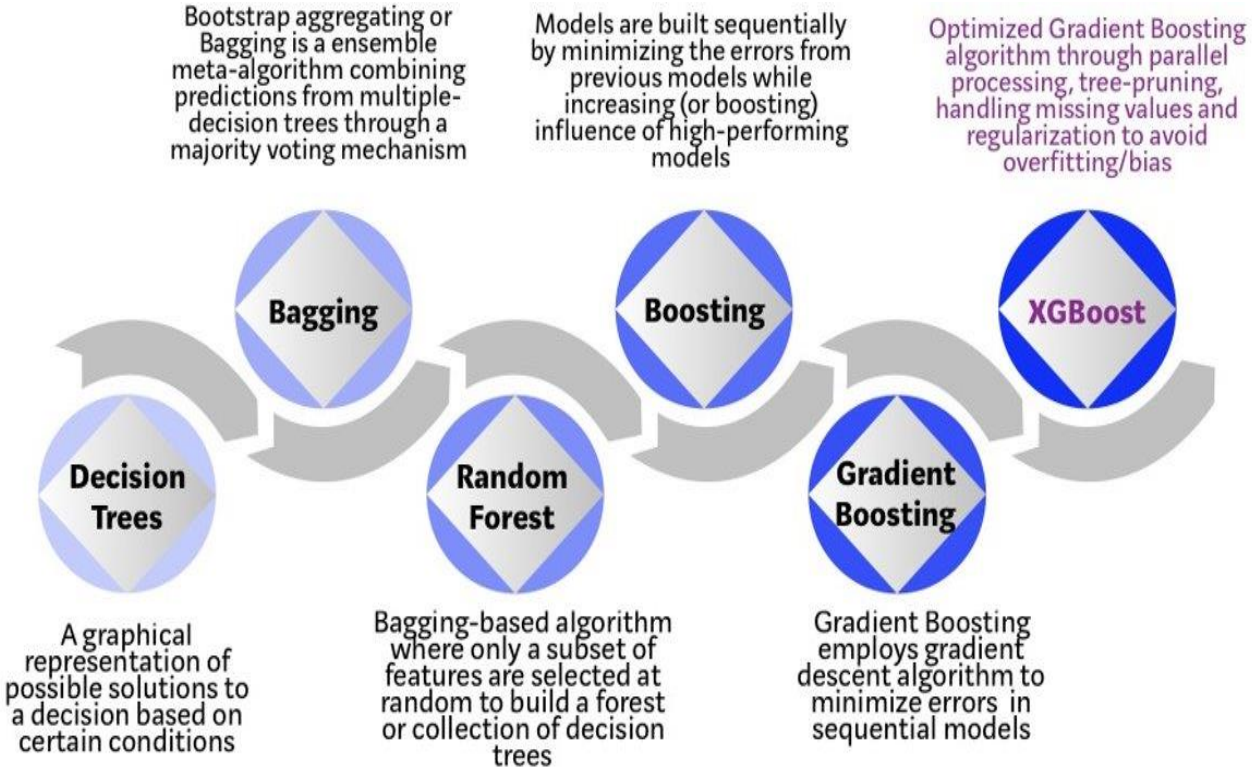

Fig. 3.1 Evolution of XGBoost [4]

3.2 XGBoost Overview

Tree is a basic data structure in computer science with a root node and subtrees, each with child and a parent node, visualized as a connection of linked nodes. Fig. 3.2 is the structure a decision tree with main node, decision node and leaf nodes.

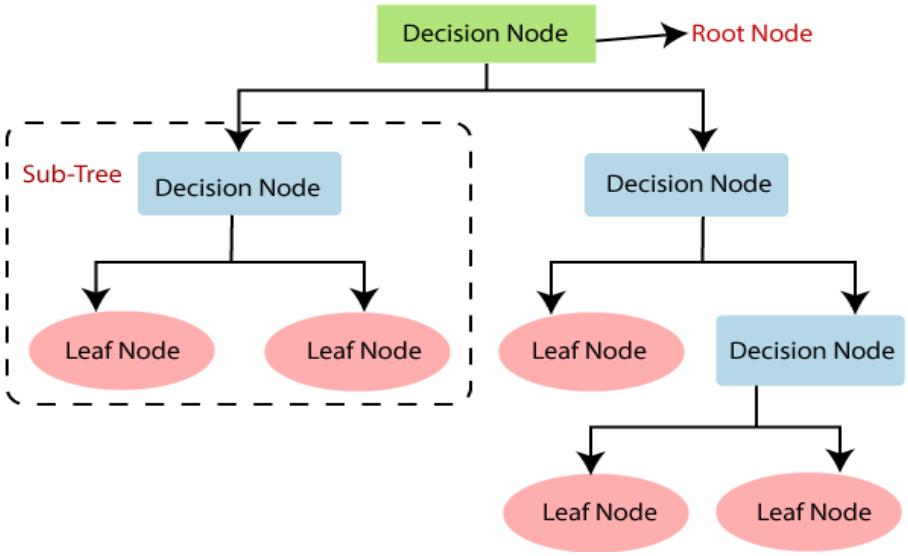

Fig. 3.2 Decision tree structure [5]

Bagging is a unique method of averaging the available data to mix the samples from the data set that is random and unpredictable. The results from all the different bagged samples are then combined using voting classifier as described in Fig. 3.3 differentiates between bagging and boosting.[15]

XGBoost improves the performance through system optimization such as:

- Parallelization
- Tree Pruning
- Hardware Optimization

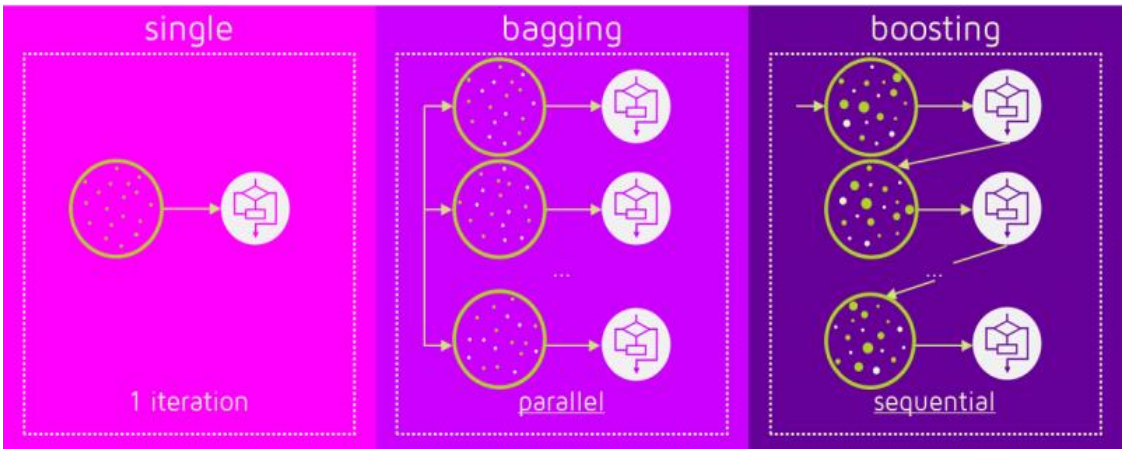

Fig. 3.3 Bagging Boosting [6]

3.3 Test-train Data splitting

For the machine learning approach, the data from 2002 to 2014 have been used. The data have been split into test and train data. The XGBoost algorithm will be modelled and trained on the train data and will be validated, accuracy be tested using the test data. The trained model will predict the values of power consumption corresponding to the time stamp in the test data and hence both predicted and actual data corresponding to particular sample is available. Fig. 3.4 illustrates the train-test data.

Fig. 3.4 Train-Test data

3.4 XGBoost Parameters

learning rate: step size shrinkage used to reduce overfitting. Range is from 0 to 1.
max depth: Quantity describes each tree is allowed to grow during boosting rounds.
subsample: Samples used for each tree. Low value leads to under fitting.
Sample by-tree: Features used per tree. High value leads to
overfitting.
N estimators: number of trees consumed to build model. XGBoost supports regularization to penalize as they become more complex and convert them to parsimonious models.
gamma: gives idea on which node will split from the giving rise to reduction in loss after the split. A high value means low splits.
alpha: first order regularization on leaf weights- large value implies more regularization.
lambda: second order regularization on leaf weights and is smoother than first order regularization. Fig. 3.5 shows the XGBoost parameters.

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0,
             importance_type='gain', learning_rate=0.1, max_delta_step=0,
             max_depth=3, min_child_weight=1, missing=None, n_estimators=1000,
             n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
             silent=None, subsample=1, verbosity=1)
```

Fig. 3.5 XGBoost Parameters

3.5 Feature Engineering

The model has been trained with the train data and the parameters given in the Fig. 3.5. Before proceeding to actual forecast it is essential to under which feature in the data contributes to the variation of the dependent variable. Feature engineering results gives us an idea of which parameters are more relevant to describe the target or predicted values. These features are listed in table 3.1.

In this project the feature that have been considered are as follows:

- Hour of day
- Day in a year
- Month of the year
- Week of the year
- Holidays
- Weekends
- T-1 values (lag by one sample)
- T-2 values (lag by two samples)

Table 3.1 Features

Datetime	Load_in_Mw	Hour	Day	Month	Year	Q	Dayofyear	Dayofmonth	Weekofyear	Drop_me	Holiday	Work	Peak	Weekend
2002-01-01 01:00:00	30393	1	1	1	2002	1	1	1	1	01-01	New Year	NonWorkTime	NonPeak	NonWeekend
2002-01-01 02:00:00	29265	2	1	1	2002	1	1	1	1	01-01	New Year	NonWorkTime	NonPeak	NonWeekend
2002-01-01 03:00:00	28357	3	1	1	2002	1	1	1	1	01-01	New Year	NonWorkTime	NonPeak	NonWeekend
2002-01-01 04:00:00	27899	4	1	1	2002	1	1	1	1	01-01	New Year	NonWorkTime	NonPeak	NonWeekend
2002-01-01 05:00:00	28057	5	1	1	2002	1	1	1	1	01-01	New Year	NonWorkTime	NonPeak	NonWeekend

The lag data of one, two and three shifts is shown in the Table 3.2. It is observed that these factor have very high correlation. This adds weight to our assumption that the present values are high dependent on the previous values.

Table 3.2 Lag data

Datetime	t-1	t-2	t-3
2002-01-01 01:00:00	NaN	NaN	NaN
2002-01-01 02:00:00	30393.0	NaN	NaN
2002-01-01 03:00:00	29265.0	30393.0	NaN
2002-01-01 04:00:00	28357.0	29265.0	30393.0
2002-01-01 05:00:00	27899.0	28357.0	29265.0
2002-01-01 06:00:00	28057.0	27899.0	28357.0
2002-01-01 07:00:00	28654.0	28057.0	27899.0
2002-01-01 08:00:00	29308.0	28654.0	28057.0
2002-01-01 09:00:00	29595.0	29308.0	28654.0
2002-01-01 10:00:00	29943.0	29595.0	29308.0
2002-01-01 11:00:00	30692.0	29943.0	29595.0
2002-01-01 12:00:00	31395.0	30692.0	29943.0

Fig. 3.6 gives us a better idea of feature engineering and where it fits in the whole modeling process. The features in their order of importance are given in Fig. 3.7 and 3.8.

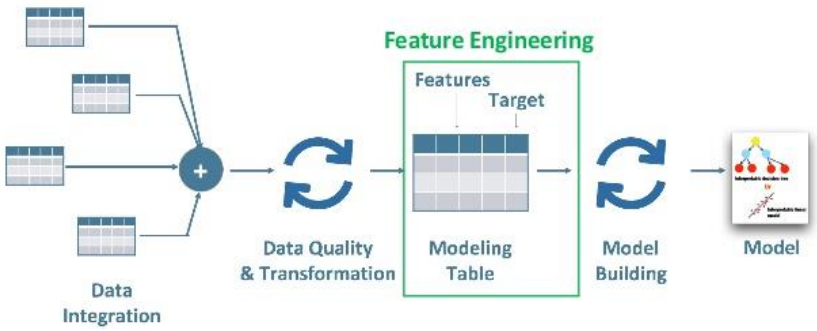

Fig. 3.6 Feature engineering

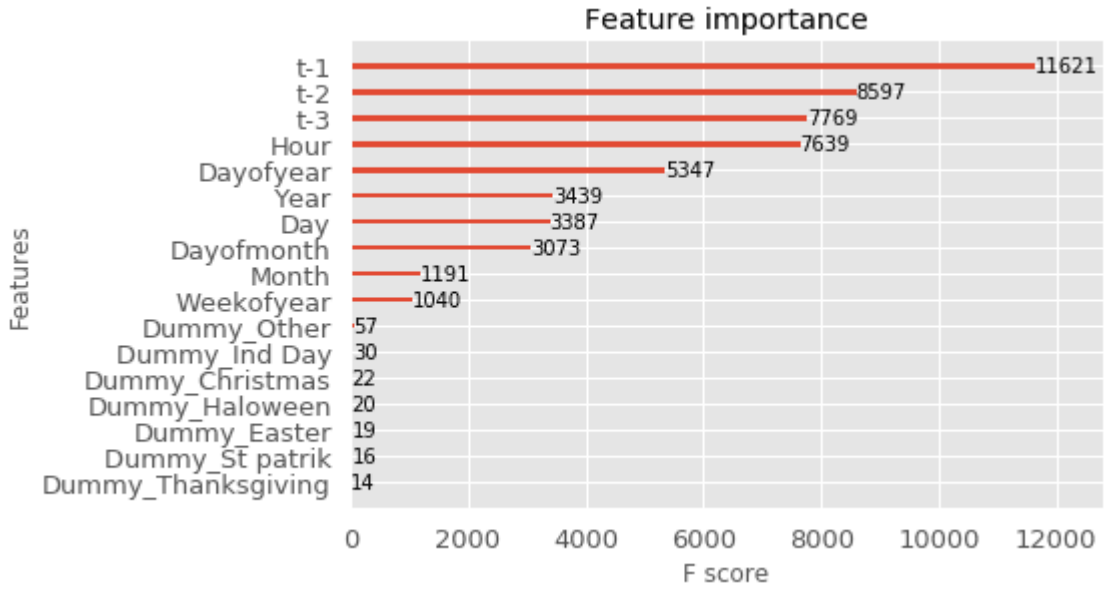

Fig. 3.7 Feature importance 1

After tuning the parameters, the result is Fig. 3.8 is obtained. It is clearly observed that t-1 has highest importance followed by hour. Other features such as Independence day, Halloween, Easter also have some importance since the consumption of power is higher during festive seasons.

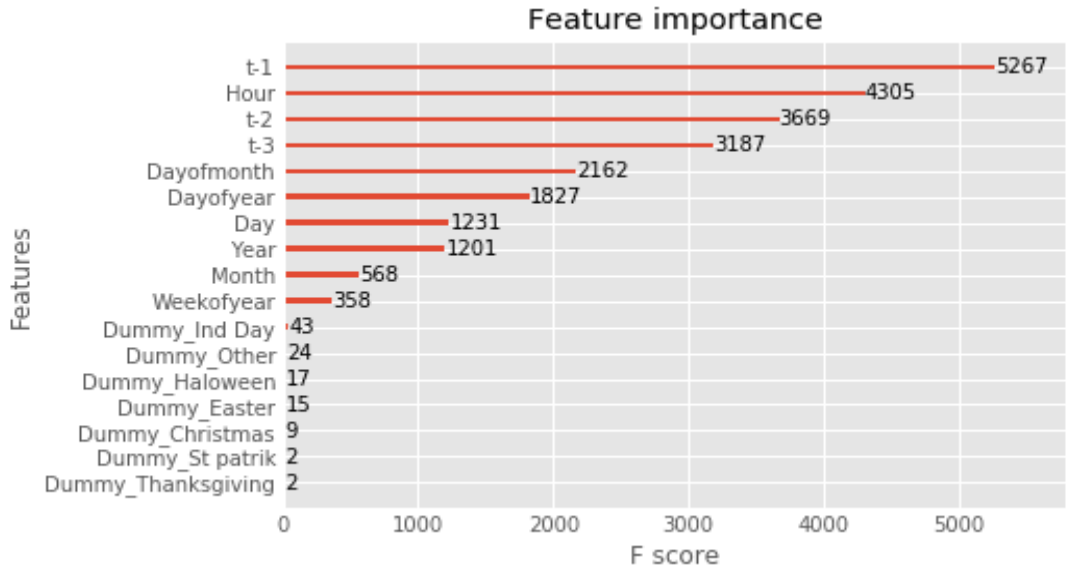

Fig. 3.8 Feature importance

3.6 Conclusion

The model has been trained and the Electric load consumption from 2015 to 2018 have been predicted. Fig. 3.9 is the graph of prediction and the actual values. The predicted values are plotted in blue and the actual values in red.

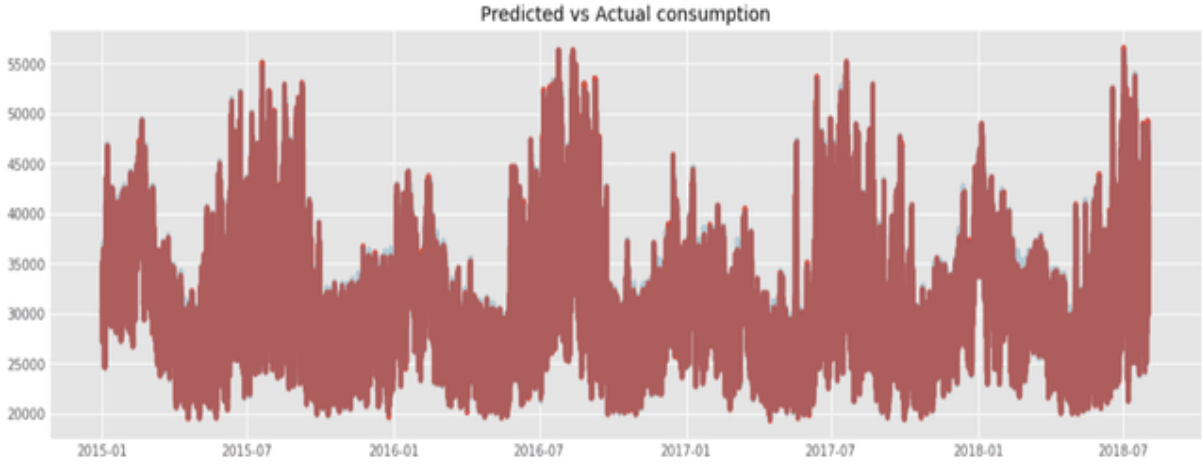

Fig. 3.9 Prediction Vs the Actual values

To understand the results much better the best and worst prediction days based on the average error aggregated to each day have been evaluated. Fig. 3.10 shows the top 10 worst days.

```
Datetime
2015-06-23 19:00:00    2969.628906
2017-03-12 20:00:00    2818.486328
2018-07-17 16:00:00    2549.058594
2016-07-18 17:00:00    2164.621094
2018-07-17 15:00:00    1836.691406
2018-02-24 19:00:00    1686.539062
2015-11-01 19:00:00    1673.048828
2016-07-14 19:00:00    1660.902344
2017-08-21 17:00:00    1660.730469
2016-02-24 19:00:00    1649.535156
Name: Difference, dtype: float64
```

Fig. 3.10 Top 10 worst predictions

Table 3.3 is the list of top ten best days; it can be observed that the samples are randomly distributed across the test data set.

Table 3.3 Top 10 best prediction

Year	Month	Dayofmonth			
2015	2	20	44694.041667	44399.632812	565.539714
2017	12	31	39016.000000	38510.500000	550.283691
		29	39392.458333	38992.015625	530.362142
2018	8	3	35486.000000	34968.410156	517.589844
2016	8	13	45185.833333	44757.187500	462.054525
2015	2	19	42278.791667	41897.511719	453.100911
2017	12	28	39963.208333	39710.207031	442.331055
2016	8	14	44427.333333	44113.332031	434.952799
2018	7	17	41080.875000	41262.718750	430.872803
2015	2	21	40918.666667	40766.257812	426.259928

The prediction vs actual graph of the test data is very huge and visualizing the whole data is not suggestive. Hence fig. 3.11 is the prediction vs actual of several days chosen randomly.

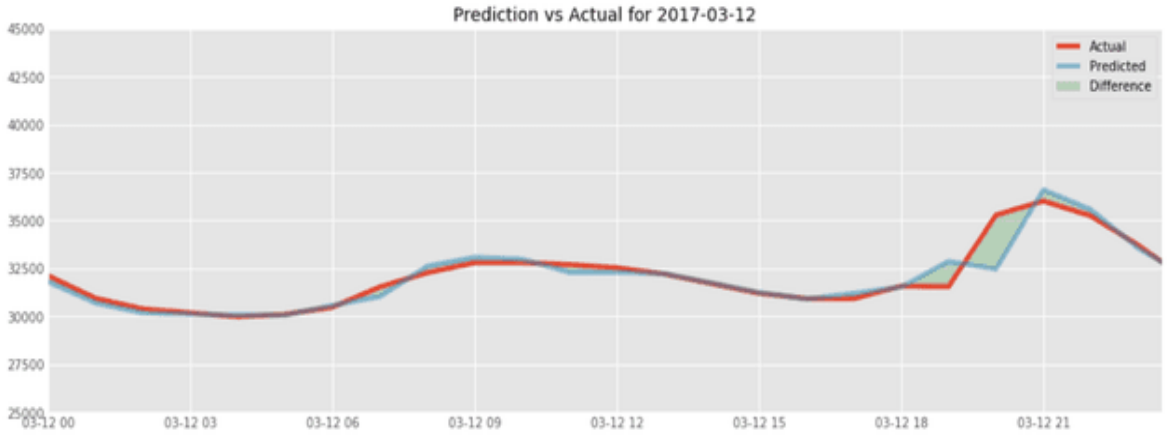

Fig. 3.11 Prediction Vs Actual for 3-12-2017

The fig. 3.12 is the prediction vs actual graph of the test data is across a week in February 2012.

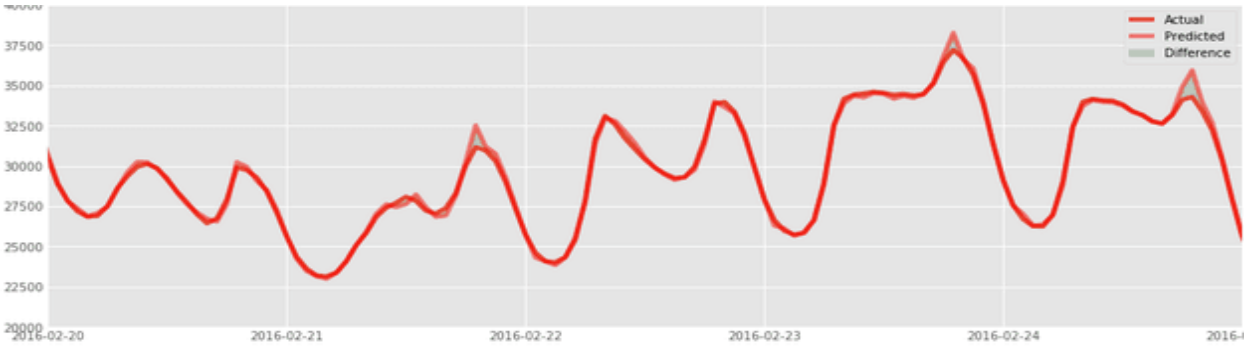

Fig. 3.12 Prediction Vs Actual for week in Feb. 2016

Fig. 3.13 is the prediction vs actual graph of 28 October 2012.

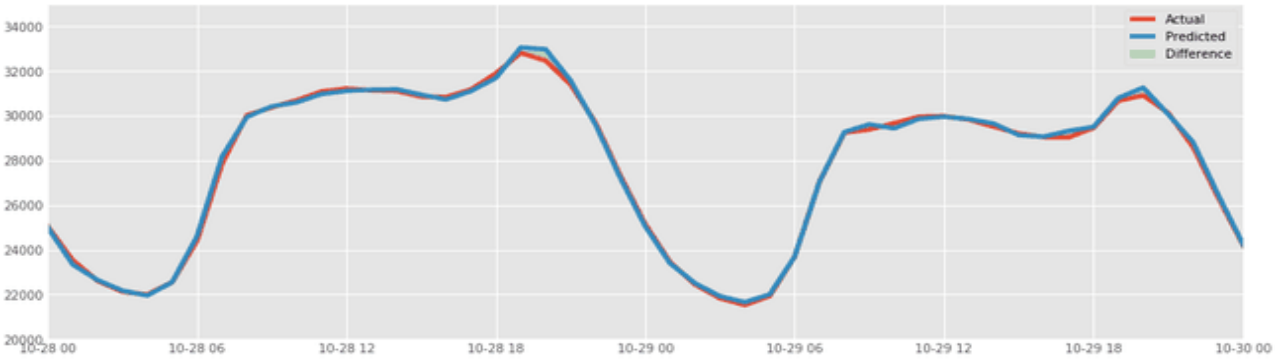

Fig. 3.13 Prediction Vs Actual 28-10-2016

CHAPTER 4

TIME SERIES FORECASTING – ARIMA

4.1 INTRODUCTION TO ARIMA

The time series data implies values that are recorded sequentially at regular intervals over a period of time. The X axis of the series is a time stamp, it can be time in hours, minutes, seconds, days, weeks, months and even years. It is described as frequency – hourly, weekly, yearly, daily and so on. As mentioned earlier the data set hourly data is used. Time series forecasting constitutes better understanding of numerous aspects of the inherent nature of the data set enabling us to be better informed to make meaningful as well as accurate forecasts.

Forecasting a time series can be classified as follows:

- Univariate Time Series Forecasting.
- Multi Variate Time Series Forecasting.

Univariate series forecasting uses only the immediate past values also known as historic data to predict its future values. Whereas Multivariate time series forecasts use exogenous variables to make the predictions. [8] For instance, in our problem, the weather data could have used as exogenous variable for our prediction.

4.2 PATTERNS AND DECOMPOSITION

Any time series data can be decomposed into multiple sub-components. There are many decomposition methods such as additive multiplicative logarithmic. They are described as additive and multiplicative. Additive can be defined as sum of trend, seasonality and white noise. While the multiplicative decomposition can be defied as the product of trend, seasonality and white noise.

The steps involved in decomposing a series are:

Step 1

Calculating the trend-cycle component using moving average method. A moving average of order m is depicted in equ. 4.1.

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}, \tag{4.1}$$

Step 2

Calculate the detrended series: original series – moving average as shown in equ. 4.2.

$$Y_t - T_t \tag{4.2}$$

Step 3

To estimate the seasonal component for each season, simply average the de-trended values for that season as shown in equ. 4.3. These seasonal component values are then adjusted to ensure that they add to zero.

$$R_T = Y_t - T_t - S_t \tag{4.3}$$

Step 4

The remainder component is calculated by subtracting the estimated seasonal and trend-cycle components.

Classical decomposition of a time series by considering the series as an additive combination of the base level, trend, seasonality and the error has been performed. The trend is inferred from the data which has an increasing or decreasing slope over the period of time series. Whereas seasonality is observed when there is a distinct repeated pattern observed between regular intervals due to seasonal factors. The stochastic reminder when actual series is differenced with trend and seasonality will give rise to residual. [9]

Fig. 4.1 is the decomposed time series. The first row is the actual time series, the second is the trend, third the seasonality and the last row is the residue.

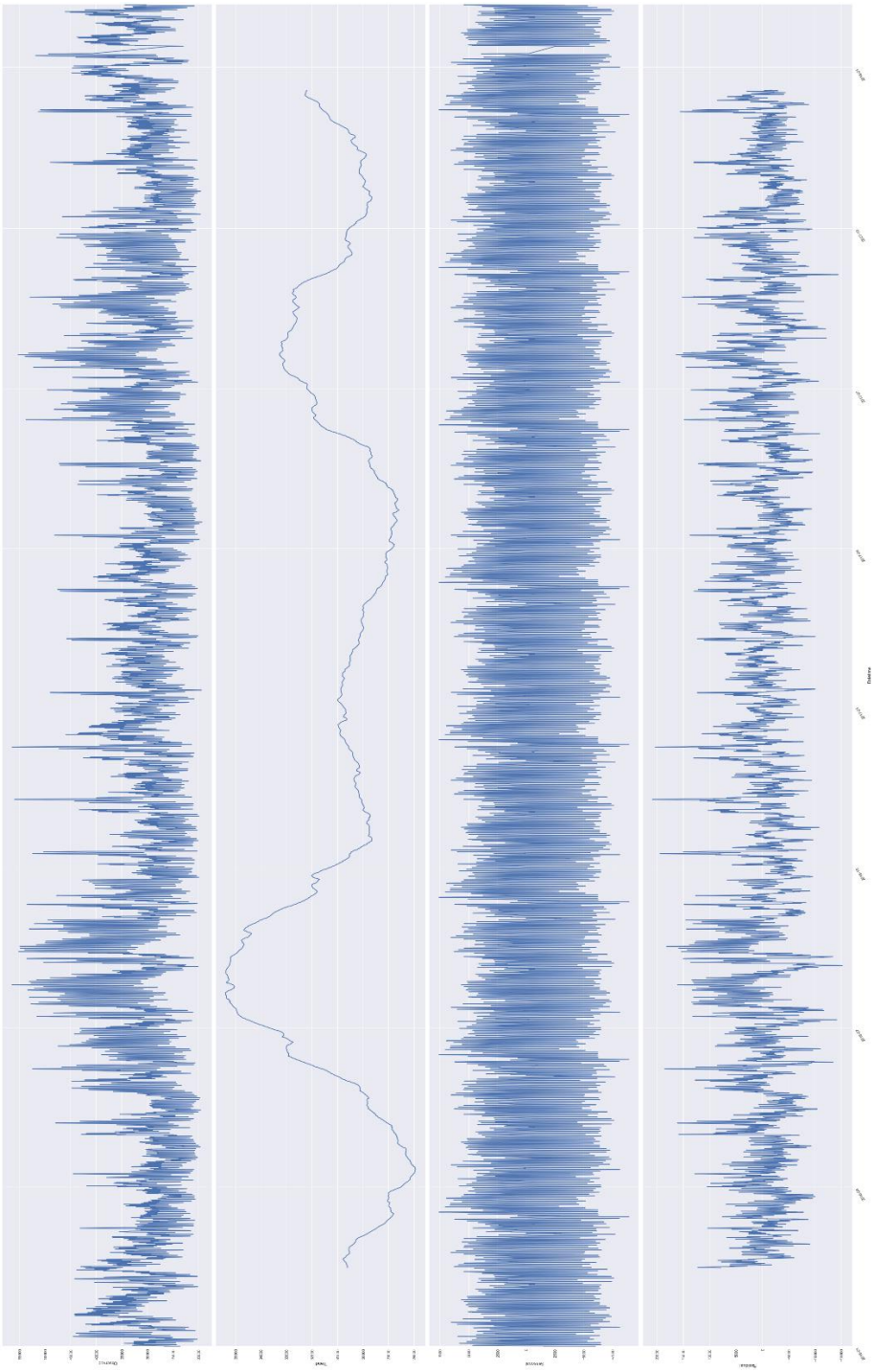

Fig. 4.1 Decomposed series

4.3 STATIONARITY

Stationarity is a property of a time series data. A stationary series is an instance in which the statistical inferences or records of the series is not a function of time. That is, the properties of the series like mean, variance and autocorrelation are constant over the period of time. Autocorrelation of the series is nothing but the correlation of the series with its previous values.

Any series can be made stationary by the following methods.

- Differencing the Series (once or more)
- Take the log of the series.
- Take the nth root of the series.
- Any Combination of the above.

There are multiple tests used in statistics to test the stationarity of a series like:

- Augmented Dickey Fuller test (ADH Test)
- Kwiatkowski-Phillips-Schmidt-Shin – KPSS test (trend stationary)
- Philips Perron test (PP Test)

Augmented Dickey Fuller test have been performed on the data. For the instance the data from January 1st 2016 to November 30th 2012 has been considered. Fig. 4.2 is the plot of the corresponding data.

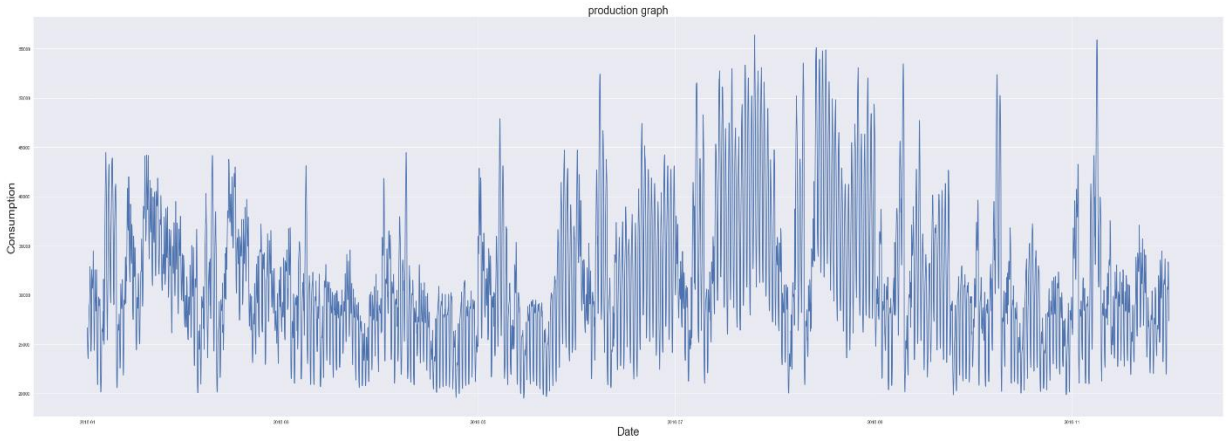

Fig. 4.2 Plot of 2016 data

Fig. 4.3 the Result of ADF test and the plot of original, moving average, and standard deviation. The line in cyan color is original data, red is moving average, black is standard deviation.

Fig. 4.3 ADF Test

The Results of dickey fuller test

Test Statistics:	-8.121675e+00
p-value:	1.151609e-12
No. of lags used:	3.400000e+01
Number of observations used:	8.005000e+03
critical value (1%):	-3.431167e+00
critical value (5%):	-2.861901e+00
critical value (10%):	-2.566962e+0

The p value is low but the ACF graph not have zero crossing hence operation such as log to remove trend has to be performed. The data has to be differenced to obtain better p value. The Pearson's correlation coefficient is a number between -1 and 1 that describes a negative or positive correlation respectively. A value of zero indicates no correlation.

The correlation for data set along rows of preceding time stamps know as lags are calculated. Since the correlation of the data set is performed with values of the same data set at lagged values, this is known as just correlation, or an autocorrelation. The plot of the autocorrelation of a data set by lag-values is knows the autocorrelation Function. This plot is often called a correlogram plot. Fig. 4.4 is the ACF and PACF plots.

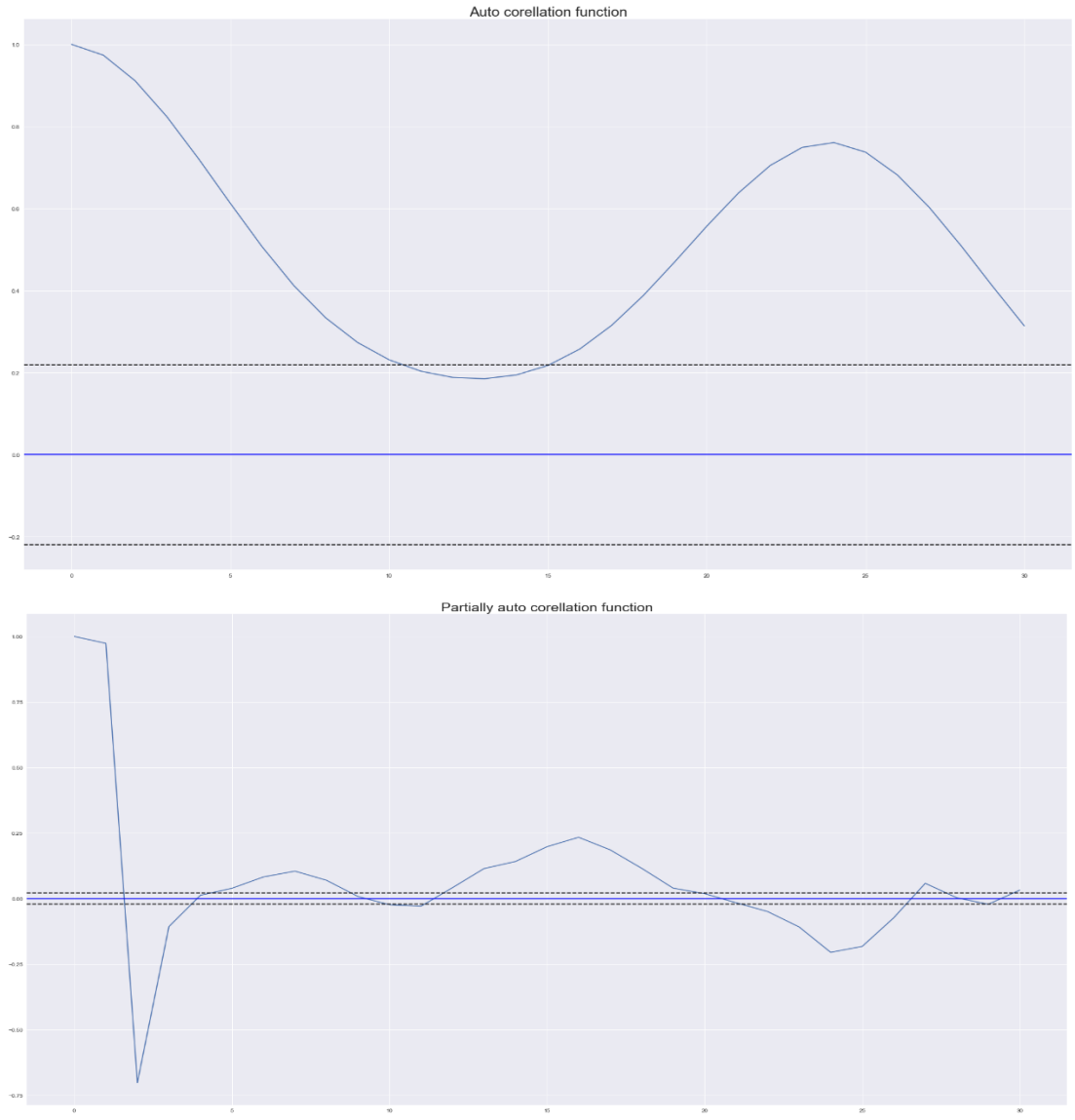

Fig. 4.4 ACF and PACF

Now, the log of the data is taken and then the difference of the rolling mean of the logged data and the actual log series is taken. ADF is performed on it. Fig. 4.5, 4.6 and 4.7 are the results. The line in cyan color is original data, red is moving average, black is standard deviation.

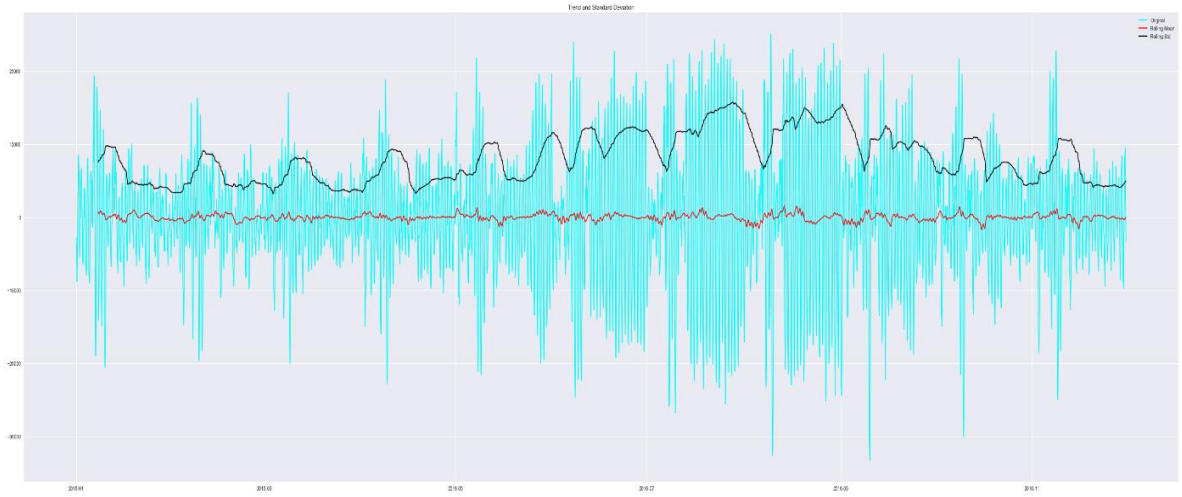

Fig. 4.5 ADF Test 1

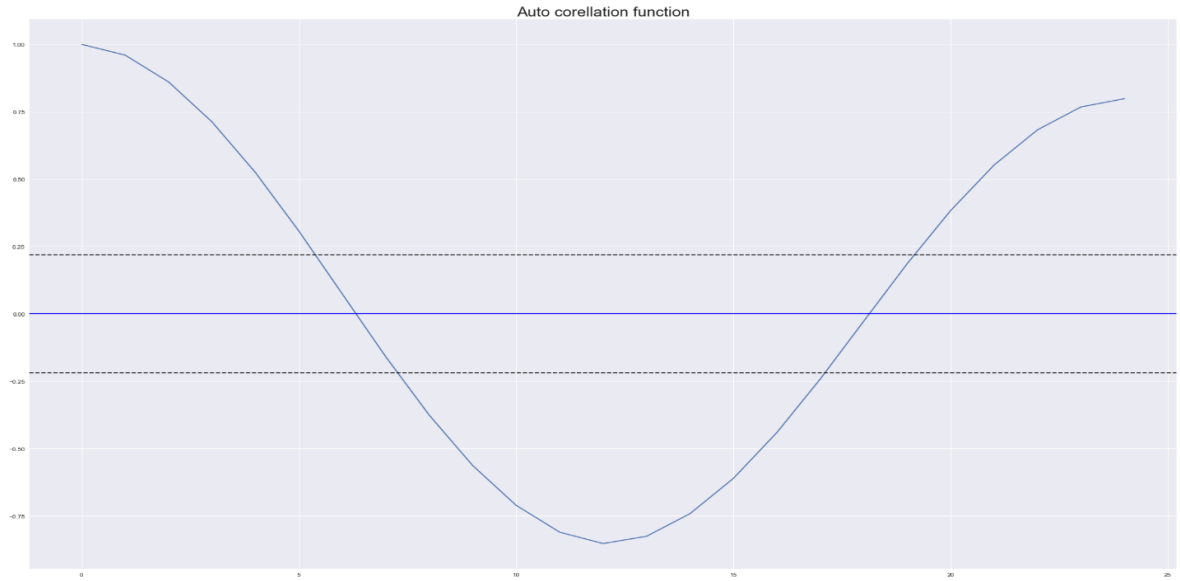

Fig. 4.6 ACF 1

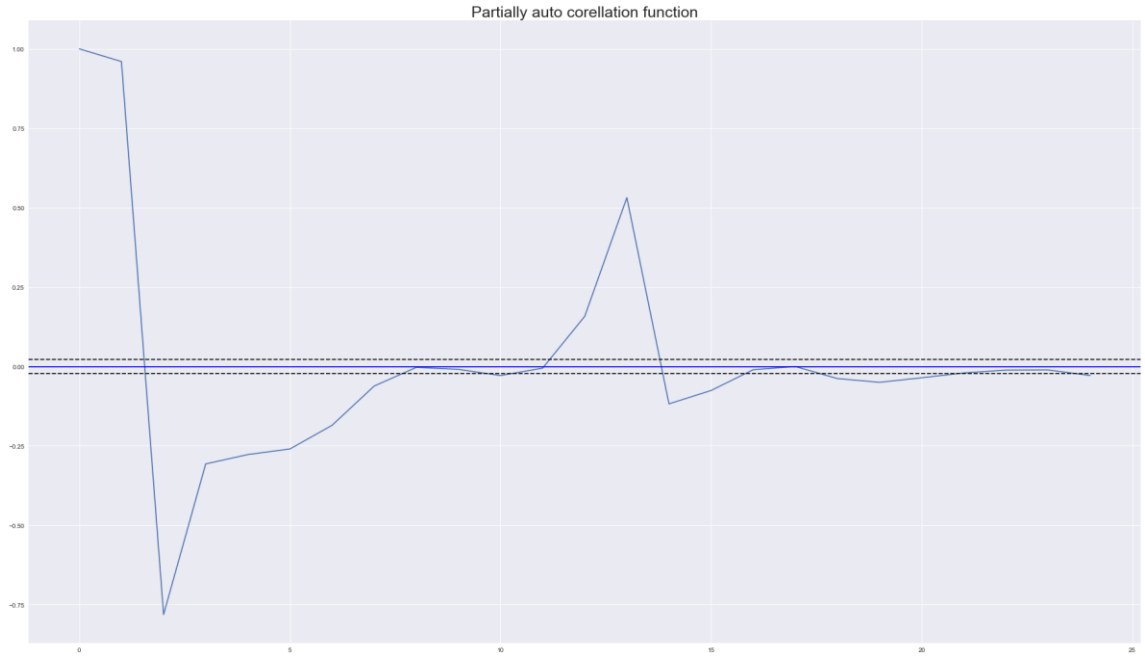

Fig. 4.7 PACF 1

Now Results of dickey fuller test

Test Statistics:	-1.345907e+01
p-value:	3.559400e-25
No. of lags used:	3.600000e+01
Number of observations used:	8.003000e+03
critical value (1%):	-3.431167e+00
critical value (5%):	-2.861901e+00
critical value (10%):	-2.566962e+00

Multiple combinations of differencing with different lags and various sequence of logs have been performed on the data and log of the data and arrived at optimal tuning. The steps followed are:

- The logarithm of the data is taken.
- The logarithm of the data is differenced with the data lag with 12 samples that is electric power consumed 12 hours past.

Fig. 4.8, 4.9 and 4.10 are the results of the ADF after performing the above mentioned transforms on the actual data, followed by ACF and PACF plots.

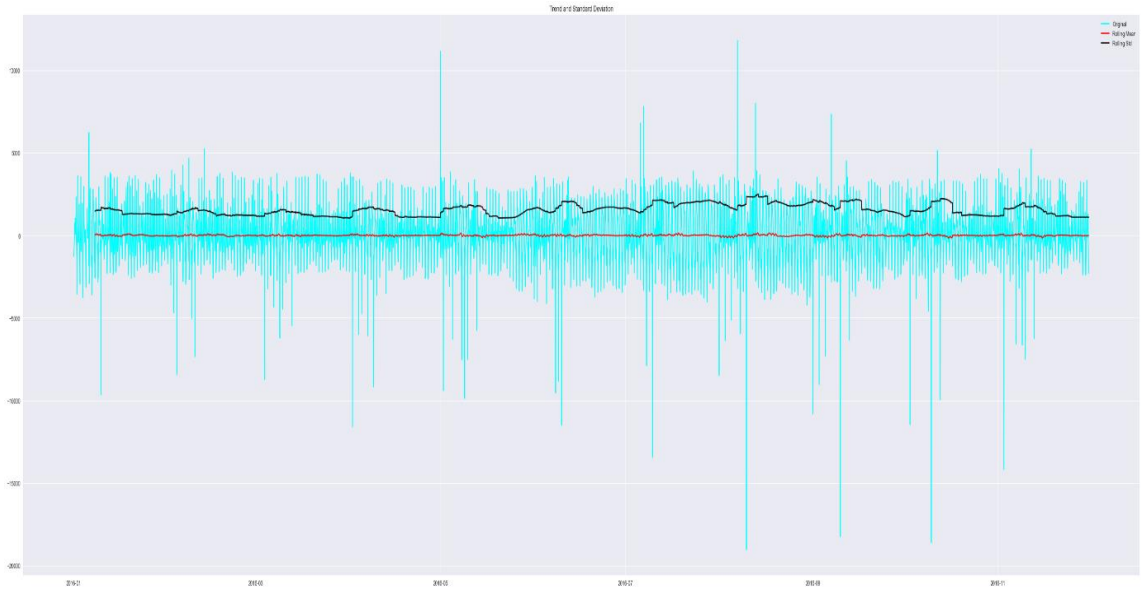

Fig. 4.8 ADF Test 2

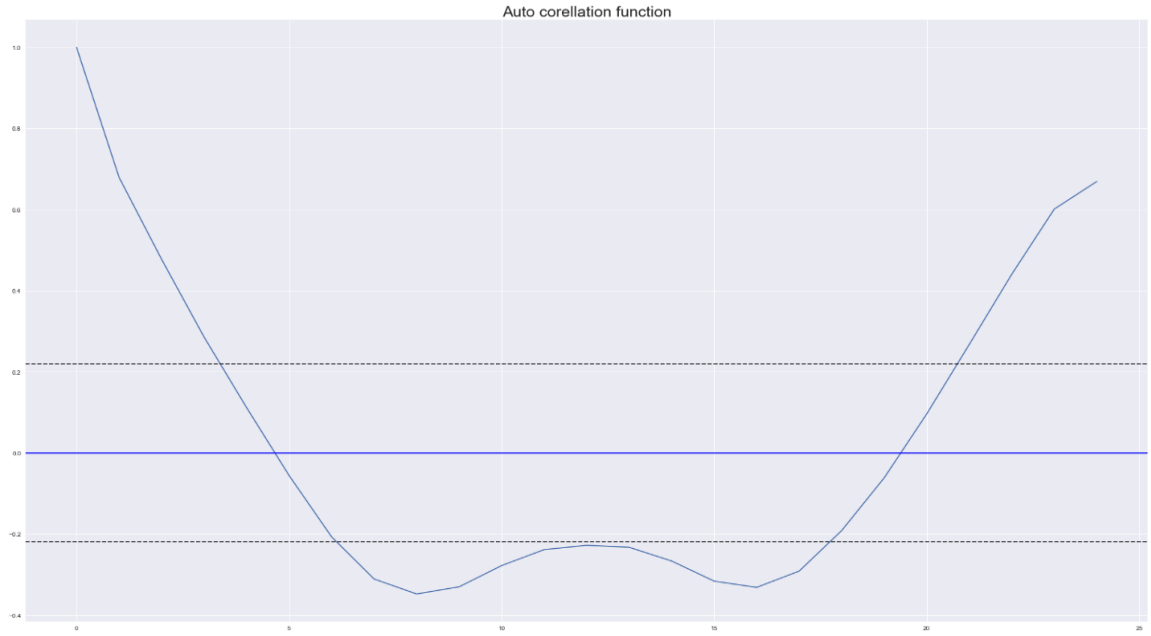

Fig. 4.9 ACF 2

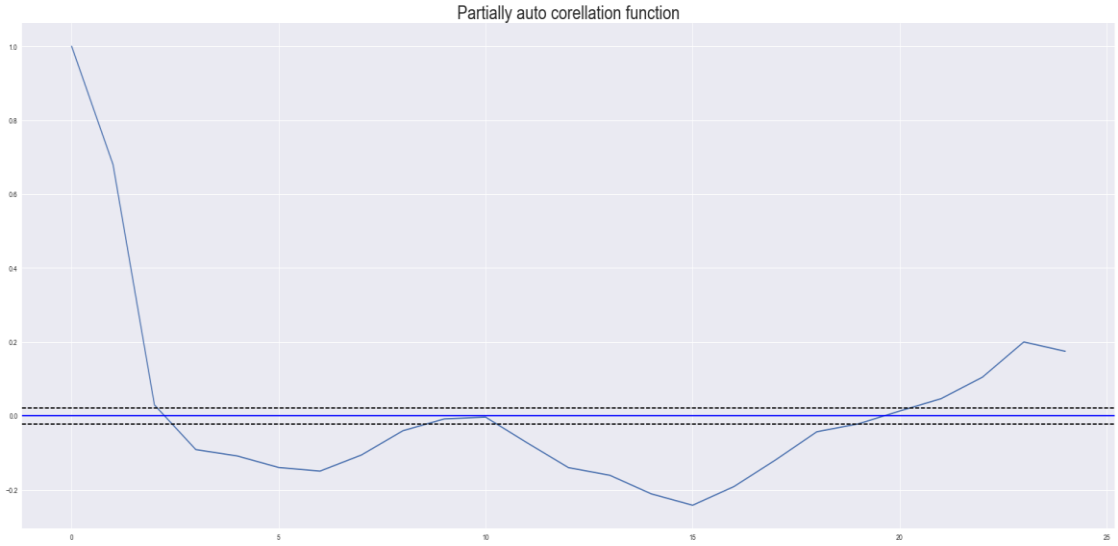

Fig. 4.10 PACF 2

Results of dickey fuller test

Test Statistics:	-1.688019e+01
p-value:	1.064376e-29
No. of lags used:	3.600000e+01
Number of observations used:	8.003000e+03
critical value (1%):	-3.431167e+00
critical value (5%):	-2.861901e+00
critical value (10%):	-2.566962e+00

Multiple combinations of values have been evaluated and the results of optimal combination is as above with p value in term of ten to the power negative 29, which is absolutely low. From the ACF and PACF plots, the optimal combination has been obtained q, d and p values for modeling ARIMA. The values of (p, d, q) that have been inferred are (2,1,4). The ARIMA (6,1,2) is also evaluated.

4.4 MODELLING ARIMA

The null hypothesis of the ADF test is that the time series is non-stationary. The probability value of the series is lower than the significant level of 0.05 then the null hypothesis is declared as rejected and implies that the series is indeed stationary. A pure Auto Regressive model is one where Y_t depends only on its lags values. That is, Y_t is a function of the 'lags of Y_t' is depicted in equ. 4.4.

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots\ldots + \beta_p y_{t-p} + \varepsilon_t \quad (4.4)$$

Likewise, a pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors. The mathematical equivalent is given in equ. 4.5.

$$y_t = \alpha + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots\ldots + \phi_p \varepsilon_{t-p} + \varepsilon_t \quad (4.5)$$

where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors from the following equations.

The final equation of an ARIMA model is given in equ. 4.6.

$$y_t = \alpha + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots\ldots + \phi_p \varepsilon_{t-p} + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots\ldots + \beta_p y_{t-p} + \varepsilon_t \quad (4.6)$$

Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags)

With the above P, Q and D values, the ARIMA model has been evaluated and the following are the summary of the ARIMA model results. The coefficients of the terms and real and imaginary root coefficients can be observed. Fig. 4.11 and 4.12 are the summary of ARIMA models.

ARIMA Model Results

Dep. Variable:	D.Load_in_Mw	No. Observations:	8039
Model:	ARIMA(6, 1, 2)	Log Likelihood	-69086.723
Method:	css-mle	S.D. of innovations	1304.999
Date:	Mon, 08 Jun 2020	AIC	138193.446
Time:	19:32:28	BIC	138263.366
Sample:	1	HQIC	138217.373

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0057	0.042	-0.136	0.891	-0.088	0.077
ar.L1.D.Load_in_Mw	0.3587	0.011	33.776	0.000	0.338	0.379
ar.L2.D.Load_in_Mw	0.9360	0.011	82.901	0.000	0.914	0.958
ar.L3.D.Load_in_Mw	-0.2661	0.015	-17.333	0.000	-0.296	-0.236
ar.L4.D.Load_in_Mw	0.0788	0.015	5.130	0.000	0.049	0.109
ar.L5.D.Load_in_Mw	-0.1017	0.011	-9.008	0.000	-0.124	-0.080
ar.L6.D.Load_in_Mw	-0.3049	0.011	-28.714	0.000	-0.326	-0.284
ma.L1.D.Load_in_Mw	3.399e-05	0.002	0.021	0.983	-0.003	0.003
ma.L2.D.Load_in_Mw	-1.0000	0.002	-609.483	0.000	-1.003	-0.997

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.9850	-0.2591j	1.0185	-0.0409
AR.2	0.9850	+0.2591j	1.0185	0.0409
AR.3	0.0730	-1.4400j	1.4419	-0.2419
AR.4	0.0730	+1.4400j	1.4419	0.2419
AR.5	-1.2247	-0.1442j	1.2332	-0.4813
AR.6	-1.2247	+0.1442j	1.2332	0.4813
MA.1	-1.0000	+0.0000j	1.0000	0.5000
MA.2	1.0000	+0.0000j	1.0000	0.0000

Fig. 4.11 ARIMA Result Summary

ARIMA Model Results

Dep. Variable:	D.Load_in_Mw	No. Observations:	8039
Model:	ARIMA(2, 1, 4)	Log Likelihood	-67827.050
Method:	css-mle	S.D. of innovations	1116.370
Date:	Mon, 15 Jun 2020	AIC	135670.099
Time:	01:34:07	BIC	135726.036
Sample:	1	HQIC	135689.241

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0009	0.013	-0.069	0.945	-0.026	0.024
ar.L1.D.Load_in_Mw	1.6257	0.010	169.078	0.000	1.607	1.645
ar.L2.D.Load_in_Mw	-0.8553	0.011	-81.266	0.000	-0.876	-0.835
ma.L1.D.Load_in_Mw	-2.0617	0.012	-178.265	0.000	-2.084	-2.039
ma.L2.D.Load_in_Mw	1.3662	0.019	72.907	0.000	1.329	1.403
ma.L3.D.Load_in_Mw	-0.0443	0.021	-2.080	0.038	-0.086	-0.003
ma.L4.D.Load_in_Mw	-0.2600	0.013	-19.294	0.000	-0.286	-0.234

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.9504	-0.5157j	1.0813	-0.0791
AR.2	0.9504	+0.5157j	1.0813	0.0791
MA.1	1.0002	-0.0000j	1.0002	-0.0000
MA.2	0.9029	-0.6904j	1.1366	-0.1039
MA.3	0.9029	+0.6904j	1.1366	0.1039
MA.4	-2.9765	-0.0000j	2.9765	-0.5000

Fig. 4.12 ARIMA Result Summary

The final step is to de-transform the data. Since various operations have been performed on the data to attain stationarity and to arrive at optimal parameters. To de-transform the same steps applied to transform are applied in exact reverse order on the forecasted data. Fig. 4.13 is the non-de-transformed result of predicted in sample data. The red is the predicted and blue line is the actual data. Note that this is not the forecast of actual data, this is the forecast of transformed data.

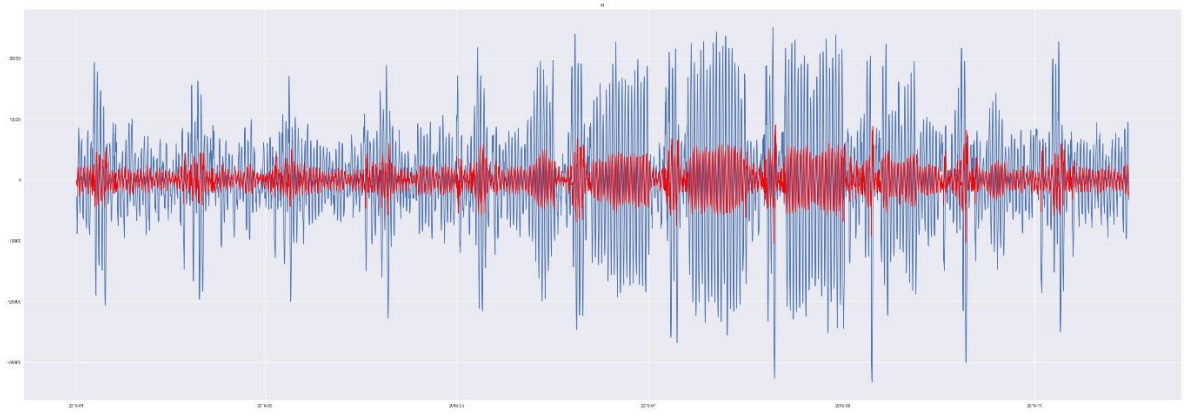

Fig. 4.13 Predicted Vs actual (yet to be transformed)

4.5 CONCLUSION

The data till 30 November 2016 has been used to evaluate the ARIMA model and one week out of the sample data has been forecasted. That is 1 December 2016 to 7 December 2012. Fig. 4.14 and 4.15 are the actual Vs forecasted plot. The green line is the forecasted values and the blue line is the actual values.

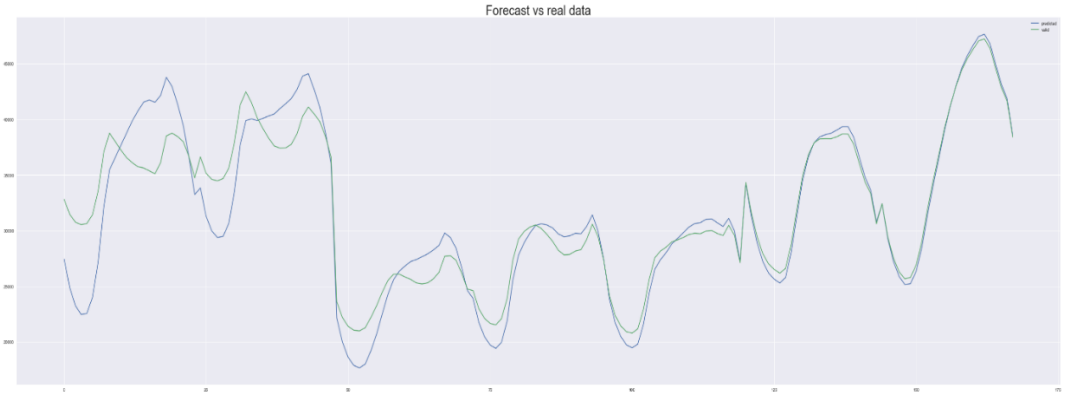

Fig. 4.14 Predicted Vs actual for out of sample (6,1,2)

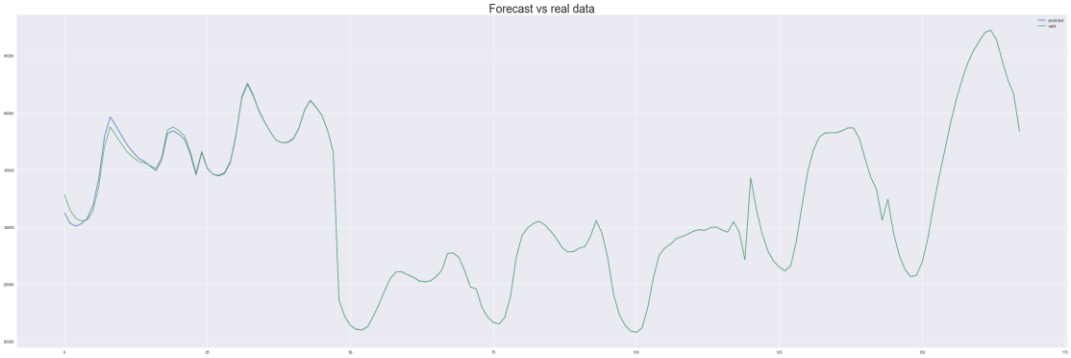

Fig. 4.15 Predicted Vs actual for out of sample (4,1,2)

The accuracy measurement is a key process to estimate how well the forecasting worked. Various accuracy measurement metric has been discussed in the previous chapters. The mean absolute percentage error (MAPE) has been evaluated.

MAPE: 2.6252560881727267

CHAPTER 5

CONCLUSION

Machine learning models and Time series analysis aid us to better understand and operate very complex power systems allowing governments, energy companies to make informed decisions and optimize losses. The primary objective to achieve more accurate forecasts for the demand of electricity has been reasonably achieved. Table 5.1 compares XGBoost and ARIMA models.

Table. 5.1 Comparison between XGBoost and ARIMA

	XGBoost	**ARIMA**
Time complexity	O (n log n)	O(n) (R language)
MAPE	9.6853	2.6252
RMSE	192.4535	17.5220
Total Computational Time	183 seconds (average)	142 seconds (6,1,2) 175 seconds (4,1,2)
Preprocessing of Data	Low	High
Data requirement	Extensive	Moderate
Further improvement	K-fold validation	SARIMAX

Both the proposed methods - XGBoost and ARIMA performed reasonably well. It has been observed that the modelling using machine learning techniques require considerable low business context that is domain knowledge where the modeling is applied. Whereas in time series analysis business context play a fairly important role. The mean absolute percentage error of the forecasts implies that the model has resulted in reasonable prediction.

Another important insight reveals the extensive use of data in forecasting. These data is generated through highly sophisticated SCADA systems across the country. In countries like India, there is strong need to implement smart metering systems to collect data and make accurate predictions.

Further cross validation and hyper tuning the parameters would have resulted in better performance of the XGBoost model. In time series forecasting more improved approach such as SARIMA would have given better performance. These advancements can be performed in further work.

REFERENCES

[1] Luis Hernández,1, Carlos Baladrón, "A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework" MDPI- Sensors (Basel), Aug 2012.

[2] [3] Farukh Abbas, Donghen Feng, "Short Term Residential Load Forecasting: An Improved Optimal Nonlinear Autoregressive (NARX) Method with Exponential Weight Decay Function." North China Electric Power University, 2018

[4] [5] [6] [15] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[7] G.J. Chen, K.K. Li, T.S. Chung, H.B. Sun, G.Q. Tang Application of an innovative combined forecasting method in power system load forecasting
Electric Power Systems Research, 59 (2001), pp. 131-137

[8] H. Hahn, S. Meyer-Nieberg, S. Pickl Electric load forecasting methods: tools for decision making European Journal of Operational Research, 199 (2009), pp. 902-907

[9] Ergogdu, E. (2007). Electricity demand analysis using co-integration and ARIMA modeling. A case study in Turkey. Munich personal RePECArchvie

[10] Forecasting: Principles and Practice Rob J Hyndman and George Athansopoulos Monash University, Australia

[11] Carlos Guestrin, Tiaqi Chen, "XGBoost: A Scalable Tree Boosting System", University of Washington.

[12] Babita Kumari Jain, Short Term Load Forecasting for Smart Power Systems", Phd. Thesis IIIT Hyderabad.

[13] https://www.machinelearningplus.com/time-series/time-series-analysis-python.

[14] Mohamed, Z. and Bodgar, P.S., "Analysis of the Logistic model for Predicting New Zealand Electricity Consumption," presented at the Electricity Engineer's Association (EEA) New Zealand 2003 Conference.