Time Series Analysis

CT/DT Number: CT20182420985
Contestant Name: Ankit Gokhroo
College Name: Jaipur Engineering College & Research Centre, Jaipur

1. Background

Problem statement : Forecast the electricity consumption of top 3 households with highest number of samples on an hourly basis based on the previous usage pattern. The major features for analysis includes household id, the plans used (standard or dynamic time of use), date, time, meter readings in Kwh and acorn groups.

The demand for electricity has been continuously increasing over the years. To understand the future consumption, a good predictive model is entailed.
Energy consumption in various households is increasing because of social development and urbanization.

Forecasting the energy consumption in households is essential for improving energy efficiency and sustainable development, and thereby reducing energy costs and environmental impact. This investigation presents a comprehensive review of machine learning (ML) techniques for forecasting energy consumption time series using actual data. Real-time data were collected from a different sources and used to evaluate the efficacy and effectiveness of statistical and ML techniques.

Well-known Artificial Intelligence Techniques were used to analyze energy consumption in single and ensemble scenarios. An in-depth review and analysis of the ‘hybrid model’ that combines forecasting and optimization techniques is presented. The predictions of model as a forecast are considered to support users or households in planning electricity management.

The purposes of this analysis a to find a model to forecast the electricity consumption in a household and to find the most suitable forecasting period whether it should be in daily, weekly, monthly, or quarterly. The time series data in our study was individual household electric power consumption. The data analysis has been performed with the ARIMA (Autoregressive Integrated Moving Average) model. The suitable forecasting methods and the most suitable forecasting period were chosen by considering the smallest value of AIC (Akaike Information Criterion) and RMSE (Root Mean Square Error), respectively. The result of the study showed that the ARIMA model was the best model for finding the most suitable forecasting period in monthly and quarterly. Then, we calculated the most suitable forecasting period and it showed that the method were suitable for short term as 28 days, 5 weeks, 6 months and 2 quarters, respectively.

2. Your Understanding

A time series is a set of statistical observations arranged is chronological order. Time series may be defined as collection of magnitudes of some variables belonging to different time periods. It is commonly used for forecasting.

Utilities of Time Series Analysis :

---

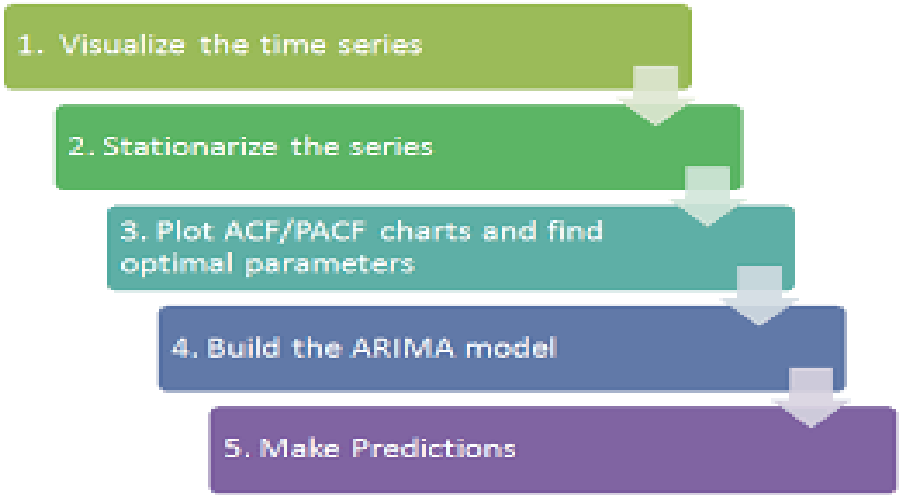
1. *It helps in understanding past behaviour and is useful for prediction of future.*

2. *It facilitates comparison.*

3. *The various components of time series are useful to study the effective change under each component.*

4. *The reasons for variation can be studied by comparing actual with expected results.*

According to problem statement , electricity consumption of various households in kwh(unit) were given, based on the previous data of electricity consumption of different households ,we need to forecast the electricity consumption of households for future or to make predictions that how much electricity household consume in future case. This forecasting helps households or others in managing the electricity consumption.



3. Scope

In Time Series analysis, based on previous data which is learn by the algorithm of our model, forecasting occurs or predictions made. This is the general approach. The ARIMA models have been extensively used for time series prediction showing encouraging results. This is theAutoregression technique which used in making forcasting for future. The basic goal of autoregression is to find a formula which forecasts each entry in a time series (except perhaps the first few) accurately from the preceding entries. It may then be reasonable to hope that the same formula can be used to forecast future entries in the series from the last few entries currently known.

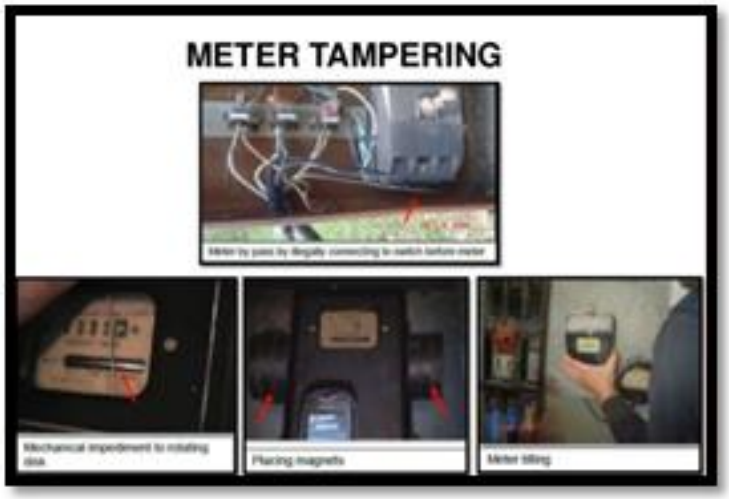
The objective of time series is to develop a mathematical model and then estimate the model to predit future patterns. Based on the forecasting of electric consumption made by the model, households will aware in using the electricity. At present also they know how much electricity they need to consume which saves electricity as they know that there is demand of more or less electricity for consumption in future case.

Scope of this time series model is very profitable to various households or others that it helps them in management of electricity consumption as it save electricity and also make avail for future use and avoid the problem of facing issues related to electricity which is a good achievement for all.

4. Out of Scope

A time series data may trend upward or downward, as many series do, or may fluctuate around a steady mean, as human body temperature does. A series may contain a single cycle, like the daily cycle of body temperature, or may contain several superimposed cycles of electricity consumption data of past. This is the this which is out of scope of the time series model and because of this it may give wrong forecasting.

These days with emerging developments in all sectors and growing demands, electricity has become priority for every individual and every organization. The basic procedure for power supply includes power generation, power transmission and power distribution to the destinations. Naturally owing to few technical faults, losses may occur due to power dissipation by some devices. These are the losses caused deliberately by human beings for the sake of illegal access to the power distribution. This is power theft..



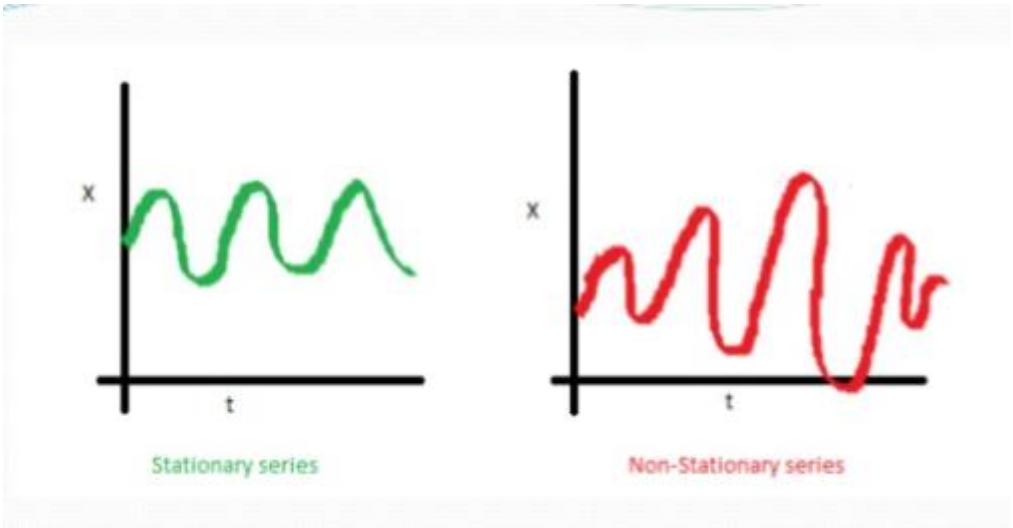
In the case of wires and cable, theft of electricity occurs due to illegally tapping to bare wires or underground cables. In the case of wires, the circuit wire is disconnected or broken from the circuit terminal block and a triple breaker inserted in the circuit.

Because of this electricity theft , meter tampering,illegal wiring there is a error in data of original electricity consumption which may also cause wrong forcasts. This should be prevented for correct forecasting.

5. Assumptions

- The data series used by arima should be stationary -by stationary it means that the properties of the series does not depend on time when it captured.

- A non stationary series should be made stationary by differencing.



6. Solution Approach

In this our main objective was to find a model to efficiently forecast the electricity consumption in a household . The suitable forecasting methods were chosen for finding the method that was suitable for short term analysis in daily, weekly, monthly, and hourly basis. The proposed forecasting time series process and the steps are shown here.

1.Data Collection :

Dataset is given of UK's households electricity consumption of past years. Dataset looks likes:

	LCLid	stdorToU	DateTime	KWh	Acorn	Acorn_grouped
0	MAC000002	Std	2012-10-12 00:30:00.0000000	0.0	ACORN-A	Affluent
1	MAC000002	Std	2012-10-12 01:00:00.0000000	0.0	ACORN-A	Affluent
2	MAC000002	Std	2012-10-12 01:30:00.0000000	0.0	ACORN-A	Affluent
3	MAC000002	Std	2012-10-12 02:00:00.0000000	0.0	ACORN-A	Affluent
4	MAC000002	Std	2012-10-12 02:30:00.0000000	0.0	ACORN-A	Affluent

Column descrption :

LCLid : Different Household Ids

KWh : Electricity Consumption in KWH(unit)

Acorn : Category of Household

Acorn_grouped : Category of Acorn

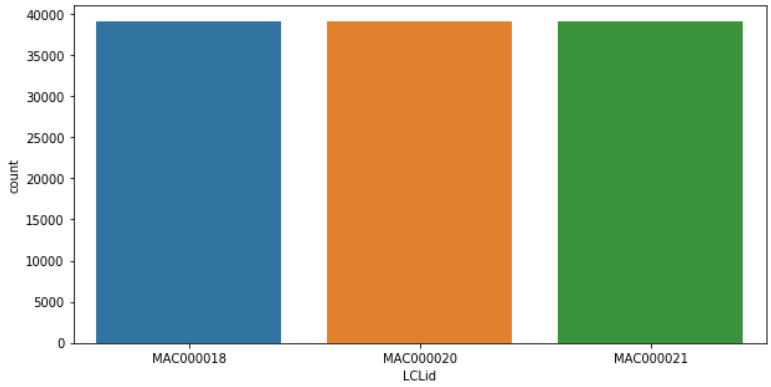
DateTime : Date and Time Column

2.Data Preprocessing :

Checked the null values by isnull() method ,the raw data are already ready for constructing forecasting model because no values are missing and the recorded time frames are appropriate.

3.Data Analysis and Visualisation:

According to problem we need forecasting of top 3 households based on samples. After analyse and visualise the data, find household Id's : 'MAC000018', 'MAC000020' and 'MAC000021' have maximum samples.



Now dataset contain only the rows having Household Id 'MAC000018', 'MAC000020' and 'MAC000021'.

After further analysis sepearted the "DateTime" column into "Date" and "Time" individually and further set "Date" column as a index. Of a dataset. By analysis , founded only "KWh" column of dataset as a important column, so droped the other columns.

Than subdivide the dataset into 3 datasets , different for each household Id. Now the datasets(3 separate) looks like :

```
Date
2011-12-07    0.303
2011-12-07    0.200
2011-12-07    0.218
2011-12-07    0.209
2011-12-07    0.210
Name: KWh, dtype: float64
```

4.Split data into train and test dataset :

Splits the dataset of each individual into train and test datasets. By the train dataset ARIMA model trained and predicts the future values of energy consumption as forecasting which are compared by the test datasets. These work individually for each household.

5.Fitting the Model :

As there are 3 different datasets for different households , trained 3 different models for forecasting of electricity consumption for different households.

Model Proposed:

Using ARIMA model, you can forecast a time series using the series past values. ARIMA models have been extensively used for time series prediction showing encouraging results.

Introduction to ARIMA Models :

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. More on that once we finish ARIMA.

So, what does the 'order of AR term' even mean? Before we go there, let's first look at the 'd' term.

term 'Auto Regressive' in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models, as you know, work best when the predictors are not correlated and are independent of each other.

The most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.

What does the p, d and q in ARIMA model mean

'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors.

And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

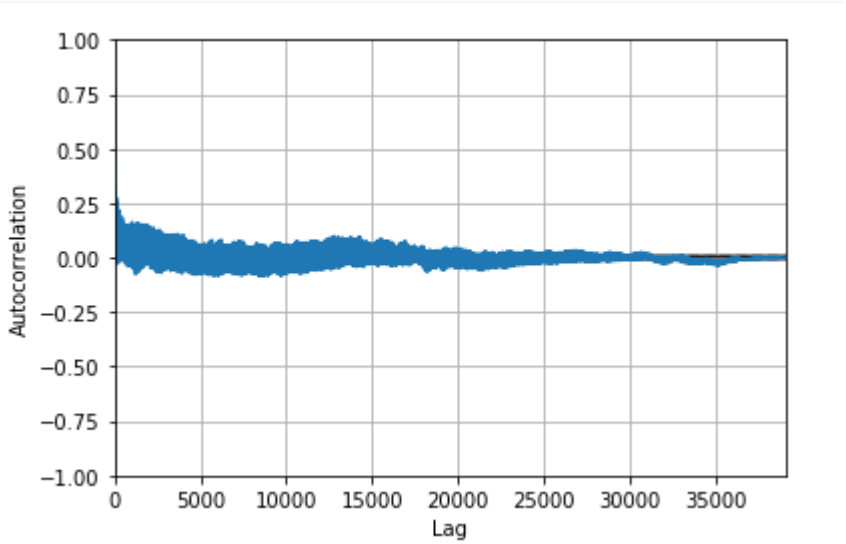
AUTOCORRELATION ANALYSIS

We can calculate the correlation for time series observations with observations with previous time steps, called lags. Because the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation, or an autocorrelation.

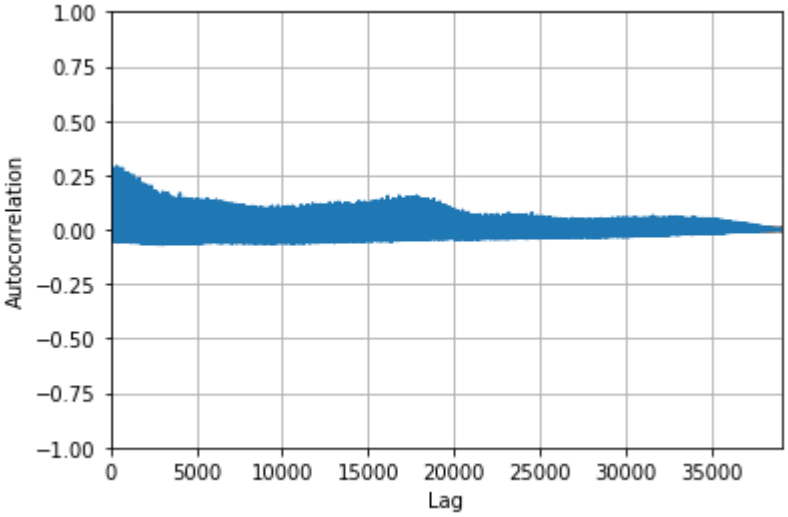
A plot of the autocorrelation of a time series by lag is called the autocorrelation function or the acronym ACF. This plot is sometimes called correlogram, or an autocorrelation plot.
A partial autocorrelation function or PACF is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

The autocorrelation for an observation and an observation at a prior time step is comprised of both the direct correlation and indirect correlations. These indirect correlations are a linear function of the correlation of the observation, with observations at intervening time steps.

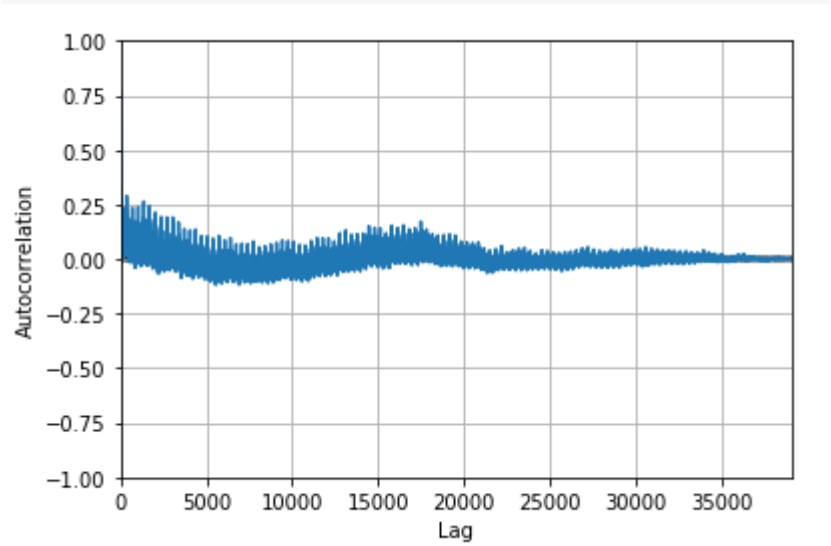
In order to calculate and plot the autocorrelation, we must convert the data into a univariate time series. Specifically, the observed daily total power consumed.



For household MAC000018



For household MAC000020



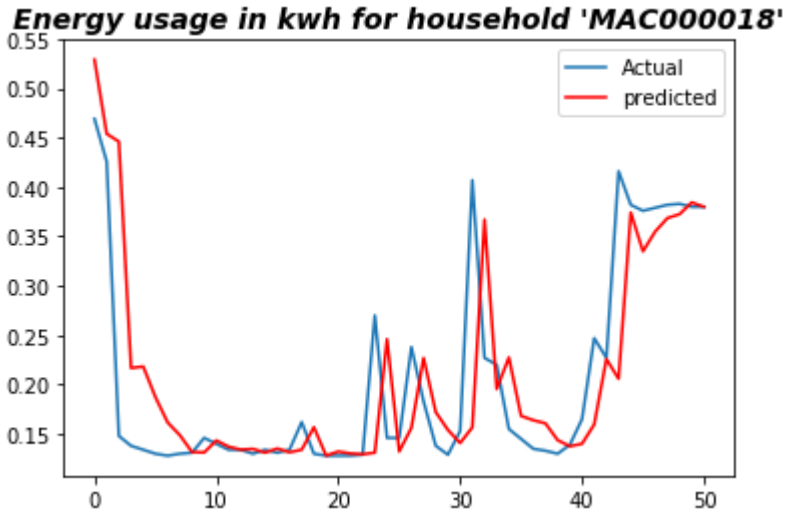
For household MAC000021

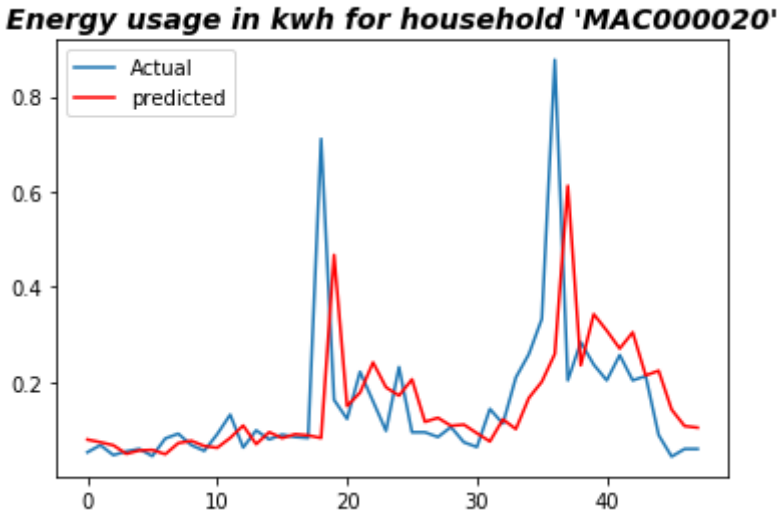
6.Build the ARIMA Model

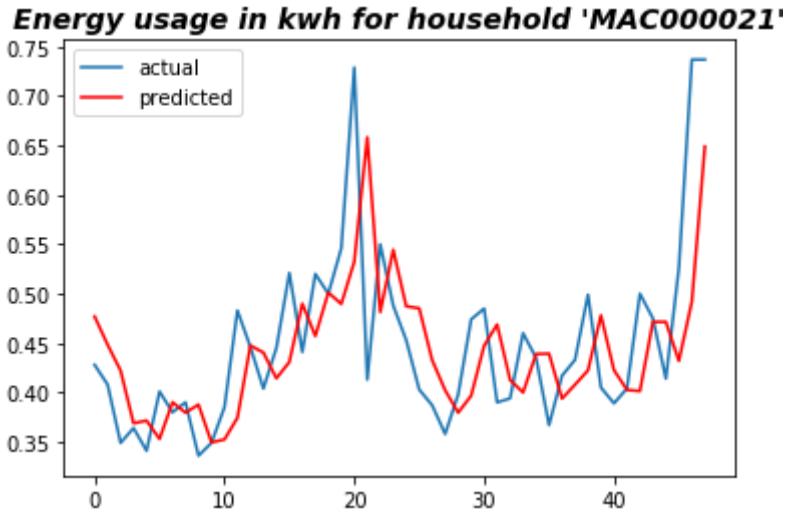
After determined the values of p, d and q, you have everything needed to fit the ARIMA model. Than just use the ARINA() implementation in statsmodel package. Arima model is trained different for the 3 households as we have 3 different datasets, hence there are 3 trained models for forecasting of energy consumptions named as “model_fit_18”, “model_fit_20” and “model_fit_21” for the 3 households.

7.Actual VS Predicted data Comparision :

Actual data from the dataset as a test data and predicted data (forecast) plotted on same curve and compared for each household.


Energy usage in kwh for household 'MAC000018'
Actual
predicted
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0
10
20
30
40
50


Energy usage in kwh for household 'MAC000020'
Actual
predicted
0.8
0.6
0.4
0.2
0
10
20
30
40



8.MODEL PERFORMANCE:

Performance for each of the 3 models for different household has been depicted by different accuracy metrices:

```
report(test_18,predictions_18)    #calling of function for household MAC000018.
```

```
mean square error: 0.005855923940638889
square root mean square error: 0.0765240089164106
mean absolute error: 0.04341732635742607
(None, None, None)
```

```
report(test_20,predictions_20)      #calling of function for household MAC000020.
```

```
mean square error: 0.024837793832365767
square root mean square error: 0.15760010733614926
mean absolute error: 0.08224576602273227
(None, None, None)
```

```
report(test_21,predictions_21)         #calling of function for household MAC000021.
```

```
mean square error: 0.006116664251133654
square root mean square error: 0.07817713381247521
mean absolute error: 0.057249801186927435
```

Summary of Arima model also depicts model performance:

ARIMA Model Results

Dep. Variable:	D.KWh	No. Observations:	39080
Model:	ARIMA(5, 1, 0)	Log Likelihood	30192.813
Method:	css-mle	S.D. of innovations	0.112
Date:	Sun, 25 Aug 2019	AIC	-60371.625
Time:	11:15:46	BIC	-60311.612
Sample:	1	HQIC	-60352.609

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.94e-06	0.000	0.009	0.993	-0.001	0.001
ar.L1.D.KWh	-0.1696	0.005	-33.604	0.000	-0.180	-0.160
ar.L2.D.KWh	-0.2454	0.005	-48.127	0.000	-0.255	-0.235
ar.L3.D.KWh	-0.1427	0.005	-27.453	0.000	-0.153	-0.133
ar.L4.D.KWh	-0.0929	0.005	-18.227	0.000	-0.103	-0.083
ar.L5.D.KWh	-0.0648	0.005	-12.833	0.000	-0.075	-0.055

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.9173	-1.2727j	1.5688	-0.1506
AR.2	0.9173	+1.2727j	1.5688	0.1506
AR.3	-0.6488	-1.6614j	1.7836	-0.3093
AR.4	-0.6488	+1.6614j	1.7836	0.3093
AR.5	-1.9717	-0.0000j	1.9717	-0.5000

For household MAC000018

ARIMA Model Results

Dep. Variable:	D.KWh	No. Observations:	39077
Model:	ARIMA(5, 1, 0)	Log Likelihood	27173.282
Method:	css-mle	S.D. of innovations	0.121
Date:	Sun, 25 Aug 2019	AIC	-54332.565
Time:	11:15:53	BIC	-54272.552
Sample:	1	HQIC	-54313.548

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.326e-06	0.000	-0.016	0.987	-0.001	0.001
ar.L1.D.KWh	-0.3920	0.005	-77.759	0.000	-0.402	-0.382
ar.L2.D.KWh	-0.3631	0.005	-67.729	0.000	-0.374	-0.353
ar.L3.D.KWh	-0.2548	0.006	-46.182	0.000	-0.266	-0.244
ar.L4.D.KWh	-0.1530	0.005	-28.539	0.000	-0.164	-0.142
ar.L5.D.KWh	-0.0843	0.005	-16.723	0.000	-0.094	-0.074

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.7371	-1.2938j	1.4890	-0.1676
AR.2	0.7371	+1.2938j	1.4890	0.1676
AR.3	-1.7991	-0.0000j	1.7991	-0.5000
AR.4	-0.7450	-1.5553j	1.7245	-0.3211
AR.5	-0.7450	+1.5553j	1.7245	0.3211

For household MAC000020

ARIMA Model Results

Dep. Variable:	D.KWh	No. Observations:	39077
Model:	ARIMA(5, 1, 0)	Log Likelihood	18582.135
Method:	css-mle	S.D. of innovations	0.150
Date:	Sun, 25 Aug 2019	AIC	-37150.269
Time:	11:15:59	BIC	-37090.256
Sample:	1	HQIC	-37131.253

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.801e-06	0.000	0.013	0.990	-0.001	0.001
ar.L1.D.KWh	-0.3326	0.005	-65.823	0.000	-0.343	-0.323
ar.L2.D.KWh	-0.1735	0.005	-32.688	0.000	-0.184	-0.163
ar.L3.D.KWh	-0.0644	0.005	-11.992	0.000	-0.075	-0.054
ar.L4.D.KWh	-0.0863	0.005	-16.250	0.000	-0.097	-0.076
ar.L5.D.KWh	-0.0469	0.005	-9.276	0.000	-0.057	-0.037

Roots

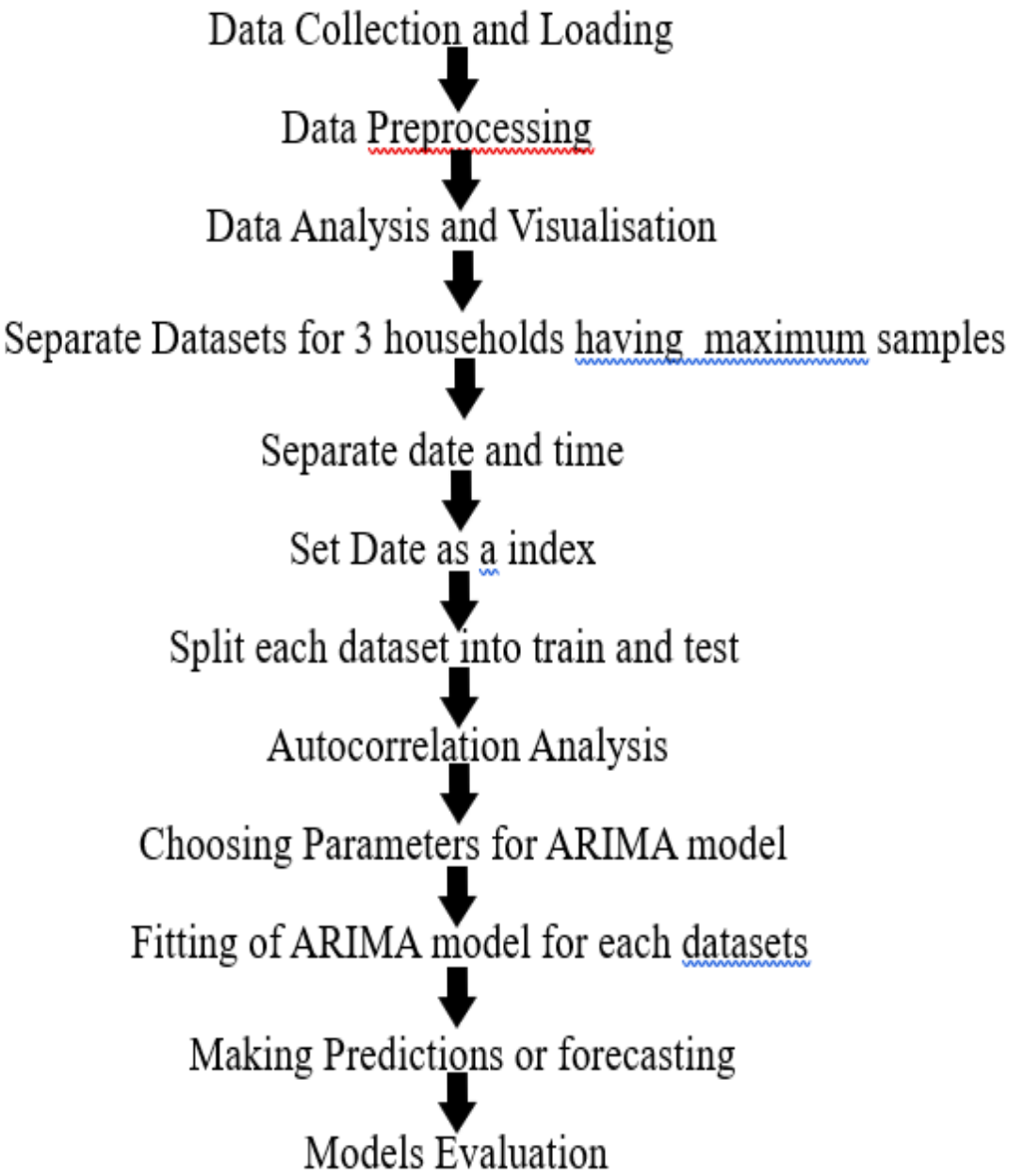
	Real	Imaginary	Modulus	Frequency
AR.1	1.0992	-1.3776j	1.7624	-0.1428
AR.2	1.0992	+1.3776j	1.7624	0.1428
AR.3	-0.9135	-1.5071j	1.7623	-0.3367
AR.4	-0.9135	+1.5071j	1.7623	0.3367
AR.5	-2.2116	-0.0000j	2.2116	-0.5000

For household MAC000021

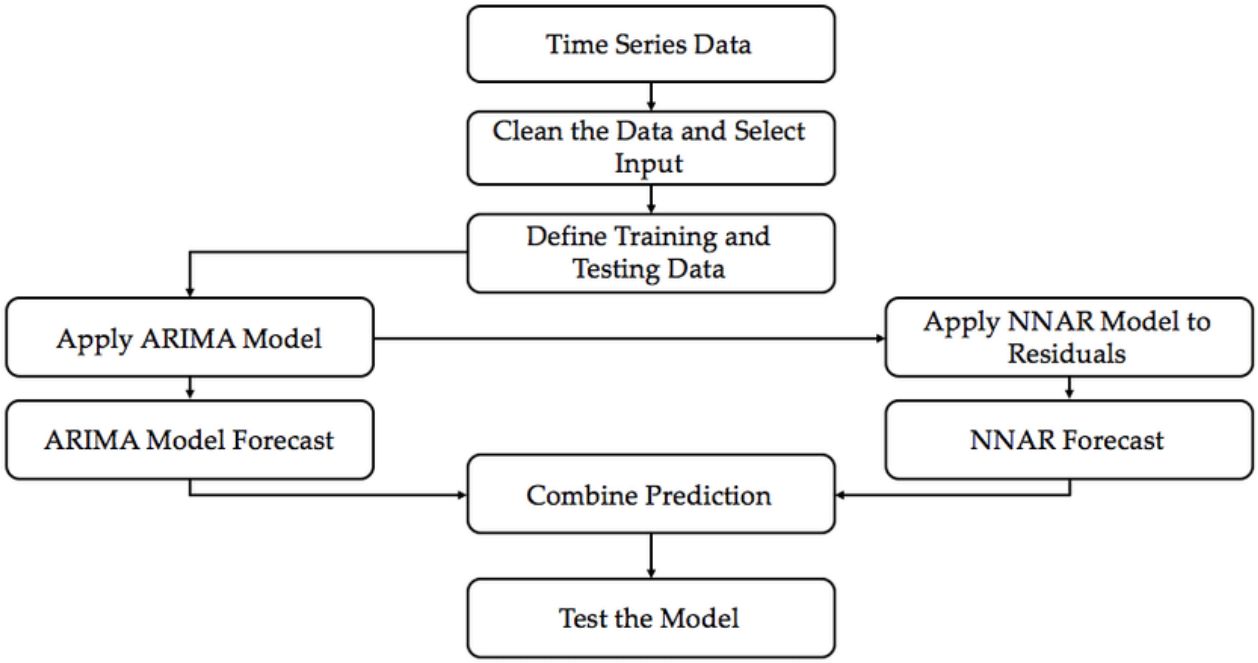
7. Implementation Framework

General Implementation Approach :

>Flow Diagram:



>Use Case Diagram :



>Actor :

Household Electric Meter

>Hardware Components :

Electric Meters and other resources for calculating and capturing the electricity consumption for data.

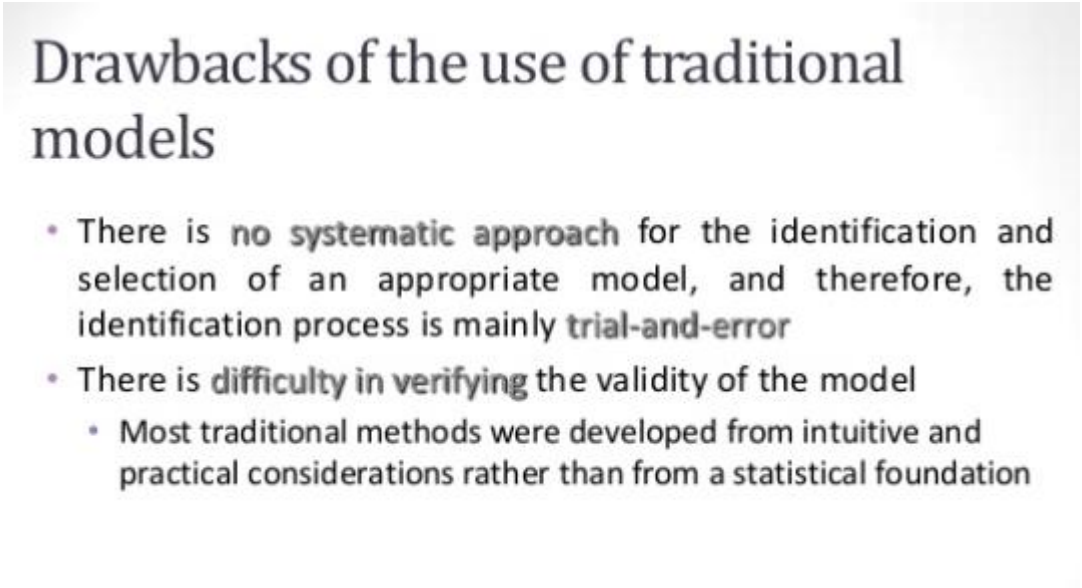
> Software Components :

Only Trained Model with user interface as a app or web app for forecasting the electricity consumptions.

8. Solution Submission

Here is the link of Submitted solution on GitHub :
https://github.com/ankitgokhroo68368/TCS-HumAIn-2019

9. Appendix



ARIMA Model

Y_t → AR filter → Integration filter → MA filter → ε_t

(long term) (stochastic trend) (short term) (white noise error)

ARIMA (2,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + b_1 \varepsilon_{t-1}$

ARIMA (3,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + b_1 \varepsilon_{t-1}$

ARIMA (1,1,0) $\Delta y_t = a_1 \Delta y_{t-1} + \varepsilon_t$, where $\Delta y_t = y_t - y_{t-1}$

ARIMA (2,1,0) $\Delta yt = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + \varepsilon_t$ where $\Delta yt = yt - yt\text{-}1$

To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q

ARIMA equations

- ARIMA(1,0,0)
  - $y_t = a_1 y_{t-1} + \varepsilon_t$
- ARIMA(2,0,0)
  - $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$
- ARIMA (2,1,1)
  - $\Delta yt = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + b_1 \varepsilon_{t-1}$ where $\Delta yt = yt - yt\text{-}1$

10. References

https://www.academia.edu/39836114/Forecasting_daily_meteorological_time_series_using_ARIMA_and_regression_models