Time Series Analysis Final Project Report

Jio Institute

Post Graduate Programme

in

Artificial Intelligence and Data Science

Group members:

Amrutha C (23PGAI0030)

Neha Thakur (23PGAI0064)

Muskan Rath (23PGAI0019)

Tanuja Tanushree(23PGAI0104)

Guided By:

Prof. Vishnuprasad Nagadevara

Saritha ma'am

CODE LINK

https://colab.research.google.com/drive/1vLIZ7laVhN_g4zxEWXgT15-1zniQ4UxV?usp=sharing#scrollTo=69383e0d

INTRODUCTION

Time series forecasting is the use of a model to predict future values based on previously observed values. Time series data is data that is collected at regular time intervals, such as daily, monthly, or yearly. Time series forecasting is used in a wide range of fields, including economics, finance, and weather forecasting. There are many different techniques that can be used for time series forecasting, including:

- Simple techniques such as mean and median, which assume that future values will be similar to the past values.
- Autoregressive Integrated Moving Average (ARIMA) models, which use a combination of autoregressive and moving average models to forecast future values.
- Seasonal decomposition, which breaks down a time series into its seasonal, trend, and residual components.

Machine learning algorithms, such as linear regression, decision trees, and neural networks, which can learn patterns in the data and make predictions based on those patterns. The choice of forecasting technique will depend on the characteristics of the time series data and the requirements of the forecasting problem.

COMPONENTS OF TIME SERIES AND THEIR EFFECTS ON FORECASTING

A time series data has three components: Seasonality, Trend and Random/Residual component.

Seasonality: Seasonality is a pattern that repeats with a fixed period of time. That fixed period of time could be daily, weekly or annually. For example: A website might be receiving more visits during weekends. In such case scenarios the seasonality will be 7 days. Another example would be in a time series of electricity demand, there may be a seasonality pattern because the demand is higher in the months of summer and lower during winter.

Trend: It refers to the underlying trend of the metrics. For example: A website which has increasing popularity would showcase a trend that is going up. In other words, it is long-term increase or decrease in the data. The trend can be either linear(for example: a straight line slope) or non-linear(example: exponential growth or decay)

Random: Also called noise, irregular or remainder, this is the residual of the original time series after the seasonal & trend series are removed. In a time series data, randomness refers to the

presence of unpredictable variations or fluctuations in the data. These fluctuations could be due to several factors, such as random errors or noise in the data collection process, external events or influences that are not captured by the model or simply inherent unpredictability of the phenomena being studied.

Effect of Seasonality, Trend and Residual on forecasting: The presence of seasonality, trend and residuals in time series data can **affect the accuracy of forecasts made using a time series model.**

a) Seasonality can make it difficult to forecast future values as the pattern of the data repeats at regular intervals and may not be easy to predict.
b) Trend can also be challenging to forecast, especially if the trend is nonlinear. If the trend is a straight line, it can make the data more predictable and if the trend is exponential,, it can make the data less predictable.
c) Since residuals are the difference between a model's observed and the predicted value. If the residual is too large, it simply indicates that the model is not suited for forecasting as it would not be accurately capture the underlying structure of the data.

Hence, it is important to remove trend and seasonality in any time series data and then apply forecasting model. This can help to reduce the influence of these components and improve the accuracy of the forecasts.

DIFFERENT METHODS OF FINDING FORECASTING RESULTS IN A TIME SERIES DATA

There are different methods that can be used for forecasting time series data. Some of the most common methods include:

1) Simple Techniques: These methods use basic statistical techniques such as the mean or median to forecast future values. These techniques are effective for data which have little to no trend or seasonality. Moving Average (MA) comes under the category of Simple Techniques where we find the mean of a set of past observations. It is also called rolling mean.
2) Autoregressive Integrated Moving Average (ARIMA) Models
3) Exponential Smoothing Models
4) Machine Learning algorithms: We can use machine learning algorithms like linear regression, decision trees and neural networks to learn patterns in the data and make forecasts based on those patterns.

a) AUTOREGRESSION: Autoregression is a statistical model that uses past values of a time series to predict future values. It is based on the idea that the current value of a time series is related to its past values. For example, in an autoregressive model of order p, the value at time t is modeled as a linear function of the p previous values:

$x(t) = c + a1x(t-1) + a2x(t-2) + ... + ap*x(t-p) + e(t)$

where x(t) is the value at time t, x(t-1), x(t-2), ..., x(t-p) are the p previous values, c is a constant term, and e(t) is an error term representing the difference between the observed and predicted values.

Autoregressive models can be used to analyze and forecast time series data and are often used in conjunction with other types of models, such as moving average models, to form the ARIMA model. The order of the autoregressive model, p, determines the number of past values that are used to predict the current value. A higher order model will be more sensitive to the past values of the time series and may be more accurate, but may also be more prone to overfitting.

b) SIMPLE MOVING AVERAGE: A simple moving average is a forecasting method that calculates the average of a set of past observations, such as k values in a time series. The average calculated is used to forecast the value at the next step.

c) EXPONENTIAL SMOOTHING: These methods use a weighted average of past values to forecast future values with more weight given to more recent values. Exponential smoothing can be used to account for trend and seasonality in the data. There are three types of exponential smoothing: a) Simple Exponential smoothing b) Double exponential smoothing(Holt's model)
c) Winter's Model

Simple exponential smoothing model is useful for series with no trend and seasonality, Holt's Method is useful for data with trend but no seasonality and Winter's method is used for data having both trend and seasonality.

The difference between moving average and smoothing is that moving averages give equal weight to past values while Smoothing gives more weight to recent observations.

d) AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS: There are three kinds of models we consider in ARIMA analysis- a) Autoregressive models b) Moving average models c) Autoregressive – Moving Average Models. ARIMA is an integration of Auto regressive (AR) and Moving Average (MA) models. The word **auto regressive** means a variable regressing on itself. It is more like a methodology rather than a model. Integration here means differentiation, which is calculating the quantity of change from one period to another. When we forecast the difference and add the difference to the last known value, this is what we call integration in case of ARIMA.
ARIMA model has three types of parameters: p (auto regressive lag) d (order of differentiation) q (Moving Average). We need to take help of ACF (Auto correlation function) and PACF (Partial Auto Correlation function) plot to get the value of p and q.

We need ARIMA models because it accounts for autocorrelation in the data, can model trends and seasonality in the data. It is primarily used to understand the past data and make future predictions for a specific time series data.

STATIONARITY IN TIME SERIES DATA

A time series is stationary if its statistical properties, such as mean and variance, do not change over time. Stationarity is an important assumption in many time series forecasting models, including the ARIMA model, as it simplifies the modeling process and can improve the accuracy of the forecasts.

Non-stationary time series data may exhibit trends, seasonality, or other patterns that change over time. These patterns can make it more difficult to model the data and can lead to forecasts that are less accurate.

There are several methods that can be used to check for stationarity in time series data, including plotting the data, calculating summary statistics, and using statistical tests such as the Augmented Dickey-Fuller test or the Kwiatkowski-Phillips-Schmidt-Shin test.

If the data is not stationary, it may be necessary to transform the data or use a different model to forecast the future values. Common techniques for making time series data stationary include differencing, which involves subtracting the value at a previous time step from the current value, and applying a mathematical function, such as taking the log of the values.

ABOUT THE PROJECT

Our project is about **Time series forecasting** of the dataset, "**Electric production**"

The dataset has been taken from the following Kaggle link:
https://www.kaggle.com/code/sercanyesiloz/electricity-production-forecasting-arima/data.

We tried studying the dataset, removed the null values, did an exploratory data analysis by plotting the dataset. Then we did a stationary check using Dickey Fuller test and removed trend to make it stationary. We decomposed time series data for further observations. For checking the seasonality, we used ACF and PACF, which will be described, in detail, in upcoming sections of the report. We have seen that there are several methods for forecasting a time series data. In this project, we are mainly using ARIMA and SARIMAX, which is an extension of ARIMA. We have then done a comparison of MSEs of ARIMA, Autoregression and Moving Average methods for both ARIMA and SARIMAX. Our observations have been recorded in the conclusion section. In every stage, we have attached the screenshots of the code and the output.

APPROACH TO THE PROJECT

PROBLEM STATEMENT

The goal of our project is to forecast electricity production using different forecasting techniques and compare their performance. We have a dataset containing monthly electricity production data for multiple years, and our task is to build models to make forecasts about future production. We will evaluate the models based on their accuracy and identify the most effective approach for forecasting electricity production.

ABOUT THE DATASET

The dataset: Electric Production, consists of 397 rows and 2 columns

MOTIVATION BEHIND CHOOSING ELECTRIC PRODUCTION DATASET

Electricity production is a phenomenon that exhibits time-dependent behavior, with production levels fluctuating over the course of months and years. Time series analysis is a powerful tool for studying electricity production data as it enables us to model and forecast future production levels based on historical data. By applying time series analysis techniques, we can identify patterns in the data, such as seasonality and trends, and use this information to make more accurate forecasts. This is important because it can help energy companies better understand how their electricity production is changing over time, and can help them make informed decisions about how to allocate resources and optimize their production processes. Additionally, analyzing electric production data over time can help policymakers and researchers better understand the factors that influence electricity production, and can inform efforts to improve energy efficiency and reduce greenhouse gas emissions.

STEPS FOLLOWED BEFORE APPLYING FORECASTING MODELS ARIMA AND SARIMAX (EXTENSION OF ARIMA)

A) IMPORTING LIBRARIES

We have imported and used the following libraries in our project:

a) pandas: A library for data manipulation and analysis, particularly for working with tabular data.

b) numpy: A library for numerical computing, including support for arrays and matrices.

c) matplotlib: A library for creating visualizations, such as plots and charts.

d) statsmodels: A python module which provides classes and function for statistical modeling, including support for time series analysis.

e) sklearn: A library for machine learning, including support for metrics and model evaluation.

f) Itertools: Itertools is a python module for working with iterators

g) statsmodels.tsa.stattools.adfuller: It is a function in statsmodels library that performs the Augmented Dickey-Fuller(ADF) test for stationarity

h) warnings: Warnings are notifications that alert the developer to potential issues or problems that may arise while running a program. These issues may not necessarily cause the program to fail, but they could indicate the use of outdated or deprecated elements, such as keywords, functions, or classes.

B) LOADING THE DATA

The original dataset has 2 columns: DATE and IPG2211A2N. The column date means the date at which the electricity production has been recorded in the string format and the column IPG2211A2N means the value of the electricity production at a particular date.

The original dataset has 2 columns: DATE and IPG2211A2N.

We read dataset as a dataframe using pandas library, renamed the columns and plotted the original data.

Fig-1: plotting the data

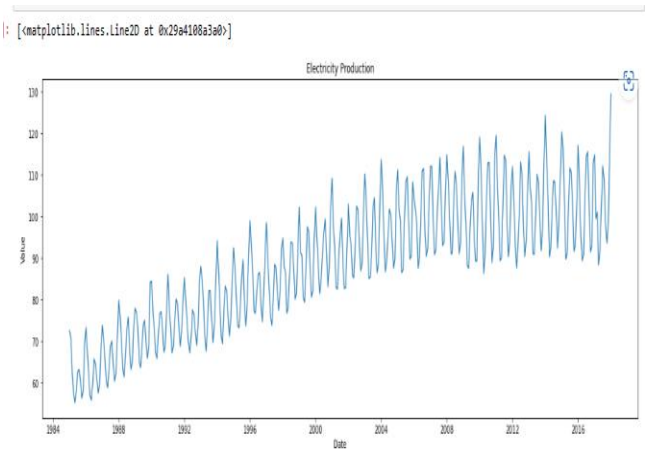


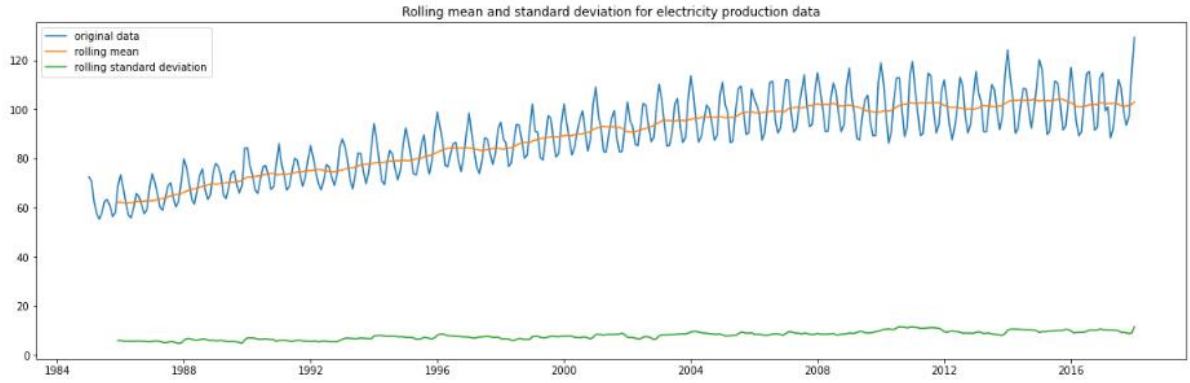
Fig-2: Renaming the columns

	Date	Value
0	1/1/1985	72.5052
1	2/1/1985	70.6720
2	3/1/1985	62.4502
3	4/1/1985	57.4714
4	5/1/1985	55.3151
...	...	...
392	9/1/2017	98.6154
393	10/1/2017	93.6137
394	11/1/2017	97.3359
395	12/1/2017	114.7212
396	1/1/2018	129.4048

397 rows × 2 columns

Observation: From the above plot, we can see that there is seasonality of 12 months and trend. There are two ways to check the stationarity of data. We have used rolling statistics and ADCF.

C) ROLLING STATISTICS



Observation:

Rolling statistics test is visual in nature. Here we have rolling mean in orange and rolling standard deviation in green. Here we can see that mean is not constant. So, we can conclude that data is **not stationary.**

D) TEST FOR STATIONARITY

First, we plotted the data to check its stationarity. Now are using rolling statistics and ADCF to check stationarity of the data.

Augmented Dickey-Fuller Test

It is a statistical test used to determine whether a time series is stationary or not. It begins with the null hypothesis that the time series is non-stationary. According to the concept of hypothesis testing, the concept of alternate hypothesis is that the time series will be non-stationary. We have denoted null hypothesis and alternate hypothesis with the notations H_0 and H_1 respectively

Null Hypothesis-

H_0: Time series is non-stationary

Alternate Hypothesis:-

H_1: Time series is stationary

If the p-value obtained from the test is less than the significance level(0.05), then we reject the null hypothesis, which means that the time series data is stationary. If p-value is greater than the significance level (0.05), then we accept the null hypothesis, which means that the time series data is non-stationary.

We used the statsmodel package as it is reliance in providing the implementation of the ADF test. The package uses adfuller() function in statsmodels.tsa.stattolls and returns the following outputs:

1) p-value
2) Test statistic value (ADF Statistic)
3) Number of lags considered for the test(n_lags)
4) Critical value cutoffs

First, we plotted the data to check its stationarity. We have carried out **Augmented Dickey-Fuller Test** to determine whether a time series is stationary or not. It begins with the null hypothesis that the time series is non-stationary.

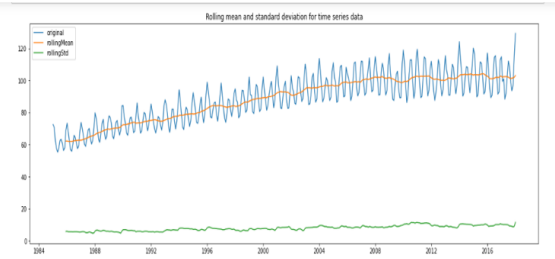


Fig-2: Graph of AD Fuller Test

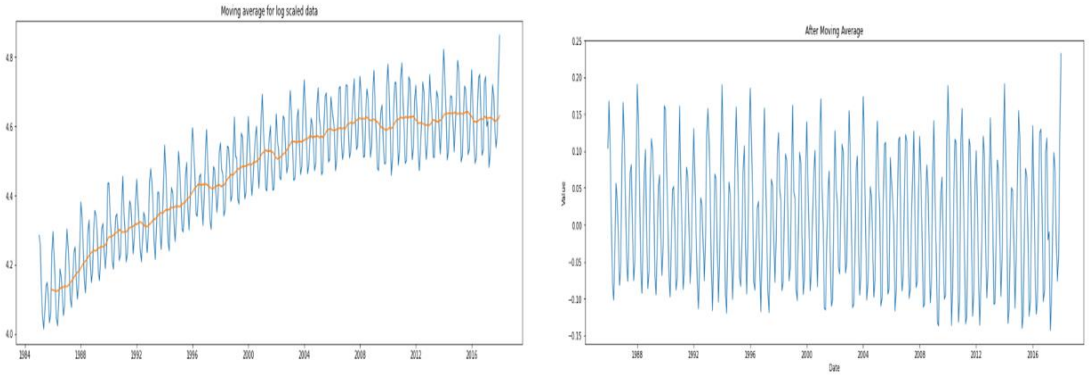
Output of AD Fuller Test:

```
ADF Statistic: -2.2569903500472366
n_lags: 0.18621469116587053
p-value: 0.18621469116587053
Critial Values:
   1%, -3.4476305904172904
Critial Values:
   5%, -2.869155980820355
Critial Values:
   10%, -2.570827146203181
```

The p value is 0.18621 which is greater than 0.05, hence the null hypothesis is not rejected and the time series data is not stationary

E) CONVERSION OF TIME SERIES DATA TO STATIONARY

Several plots for transformations: i) Log Transformation ii) Removing Trend with Moving Average



Observation: The data is stationary now.

F) EXPONENTIAL DECAY TRANSFORMATION

Exponential decay transformation is a type of transformation that is often used on time series data to remove a trend or pattern that exhibits exponential decay.

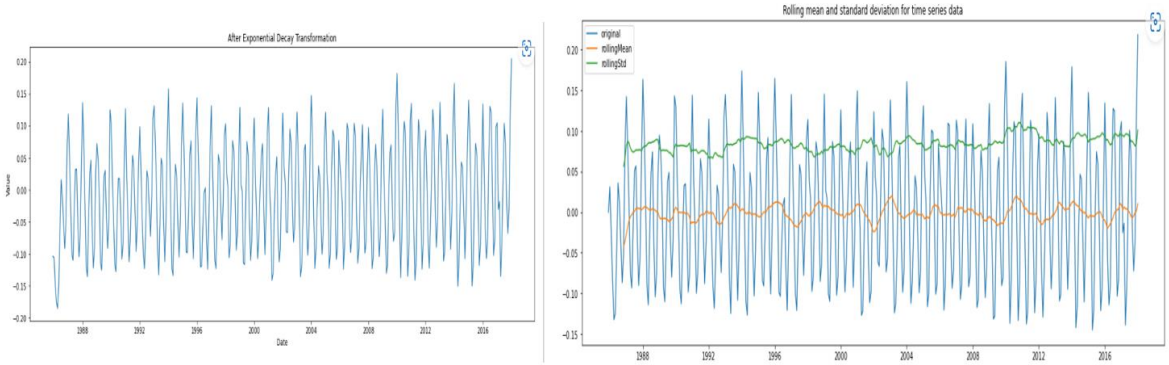


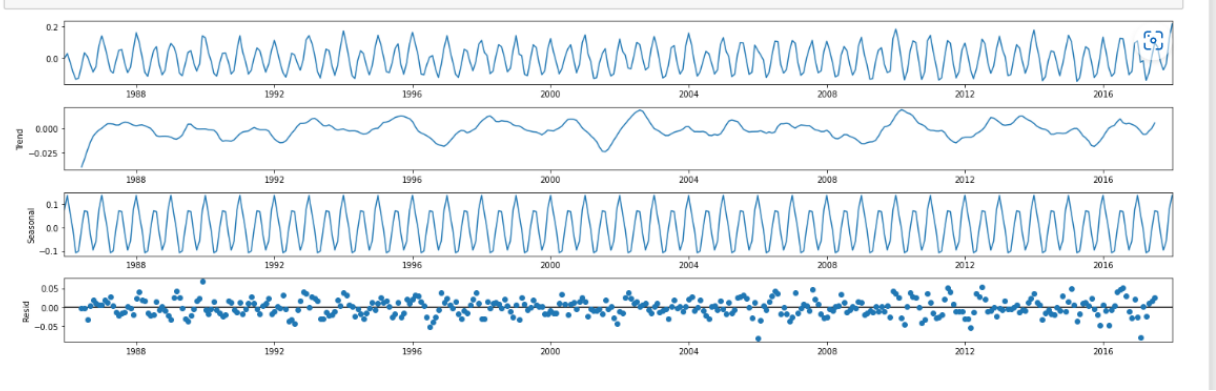
Fig: Rolling mean and standard deviation for time series data

G) DECOMPOSITION

We want to see the components of the time series.

The time series data was decomposed into the components: trend, seasonal and residual and plots were shown for each component separately.

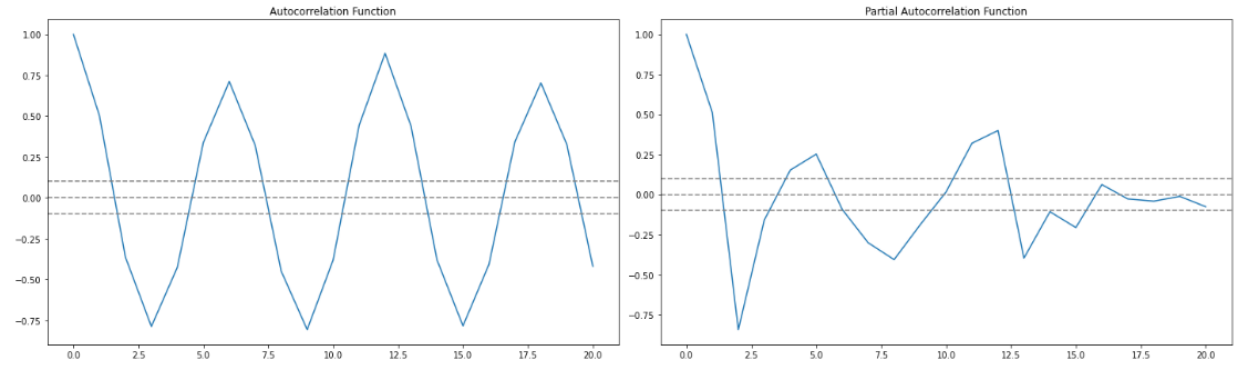
Fig: Graph of decomposition



H) ACF and PACF for checking seasonality

In Autocorrelation Function (ACF), we correlate time series with a lagged version of itself. In Partial Autocorrelation Function (PACF), we get a partial autocorrelation of a stationary time series with its own lagged values.

Fig: PACF and ACF Graph



Observation: By looking at both the ACF and PACF graph, we selected the values of p and q for auto regression and moving average model respectively of p,q and d. (p,d,q) = (2,1,2).

APPLYING FORECASTING MODELS

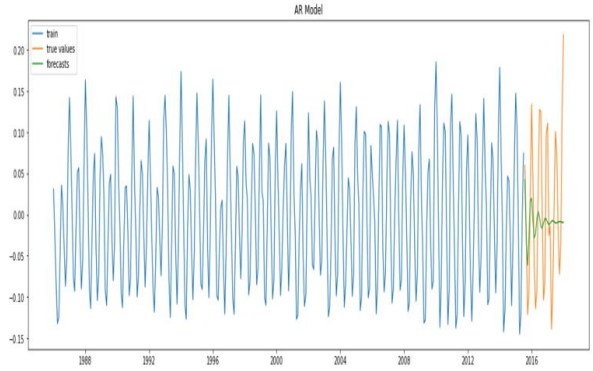
The following steps have been done for all the models:

1) Building the model, dividing the model into train and test data
2) Predicting and forecasting values

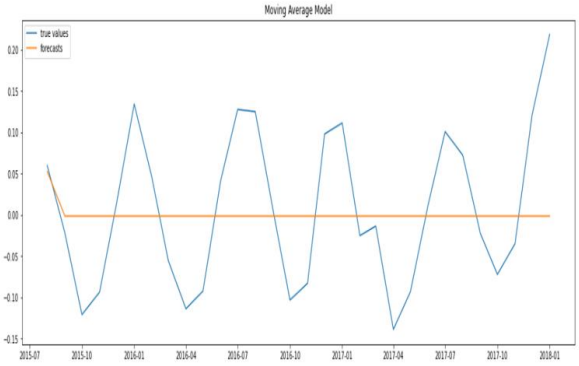
At last, we have compared MSE values of all the models.

The forecasted values for all the models have been shown in the below graphical representation:

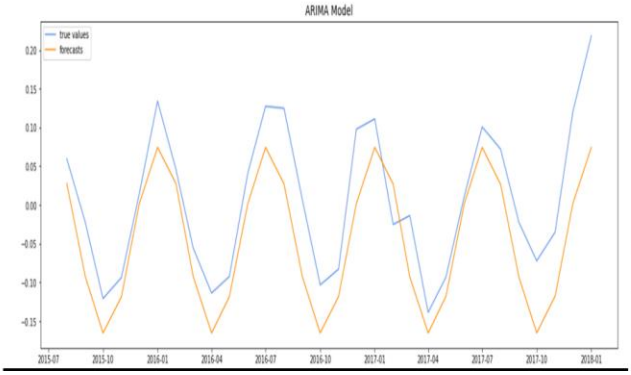
A) AUTOREGRESSION MODEL



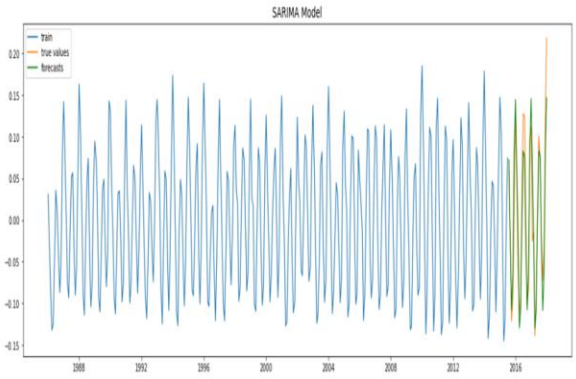
B) MOVING AVERAGE



C) ARIMA



D) SARIMA



COMPARING MSE OF ARIMA, AUTOREGRESSION AND MOVING AVERAGE

Model	MSE
ARIMA	0.004172
Autoregression	0.007904
Moving Average	0.008520

We can see that MSE of ARIMA is the least, hence, it is a better model

D) SEASONAL ARIMA (SARIMA)

SARIMA is an acronym for Seasonal Autoregressive Integrated Moving Average, and it is a type of time series model that is used to forecast data that exhibits seasonality. It is an extension of the ARIMA model, which is used to model non-seasonal data, and adds additional terms to model the seasonality in the data. SARIMA models are defined by three terms: p, d, and q, which control the autoregressive, differencing, and moving average aspects of the model, respectively. The seasonal terms are defined by three additional terms: P, D, and Q, which control the seasonal autoregressive, seasonal differencing, and seasonal moving average aspects of the model, respectively.

Finding the appropriate combination of parameters for a SARIMA model is an important step in time series modeling, as the choice of parameters can have a significant impact on the performance of the model. So, we tried a range of different combinations of parameters and evaluate their performance using a metric such as mean squared error (MSE) or root mean squared error (RMSE). The combination of parameters that results in the lowest error was the one that was likely to provide the best fit for our data.

We have plotted residual plots as below and tried several possible combinations and finally used the combination (1,0,0) x (1,0,1,12). We found the summary:

Fig: Summary

SARIMAX Results

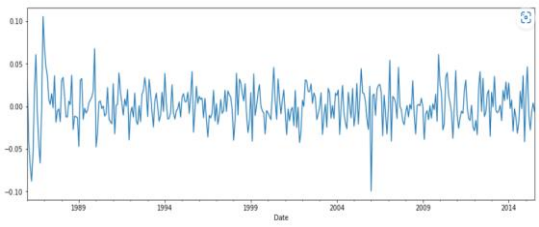
Dep. Variable:	Value	No. Observations:	355
Model:	SARIMAX(1, 0, 0)x(1, 0, [1], 12)	Log Likelihood	832.038
Date:	Sat, 07 Jan 2023	AIC	-1656.076
Time:	15:34:12	BIC	-1640.736
Sample:	01-01-1986 - 07-01-2015	HQIC	-1649.965
Covariance Type:	opg		

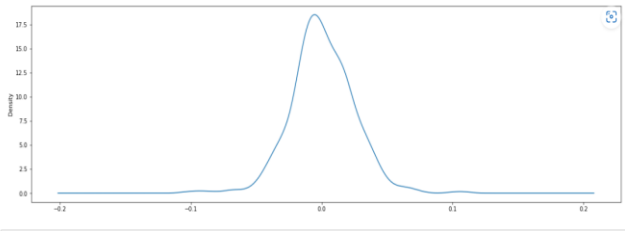
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5305	0.045	11.851	0.000	0.443	0.618
ar.S.L12	1.0061	0.003	351.132	0.000	1.000	1.012
ma.S.L12	-0.8687	0.037	-23.520	0.000	-0.941	-0.796
sigma2	0.0004	2.76e-05	15.576	0.000	0.000	0.000

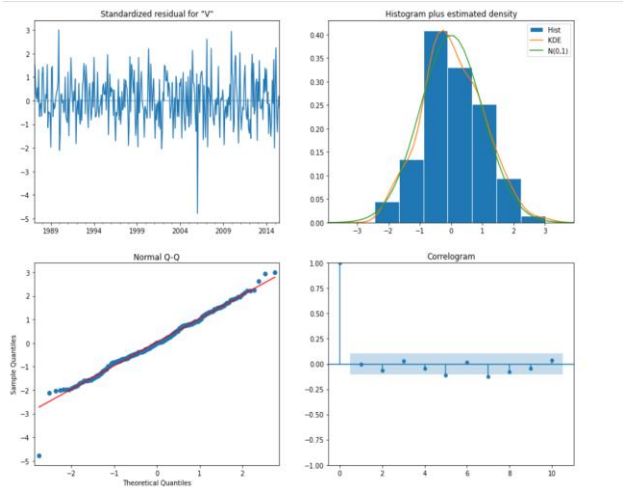
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	17.26
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	1.34	Skew:	-0.11
Prob(H) (two-sided):	0.12	Kurtosis:	4.08

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Residual plot







Model	MSE
SARIMA	0.001167
ARIMA	0.004172
Autoregression	0.007904
Moving Average	0.008520

Forecasting using SARIMA

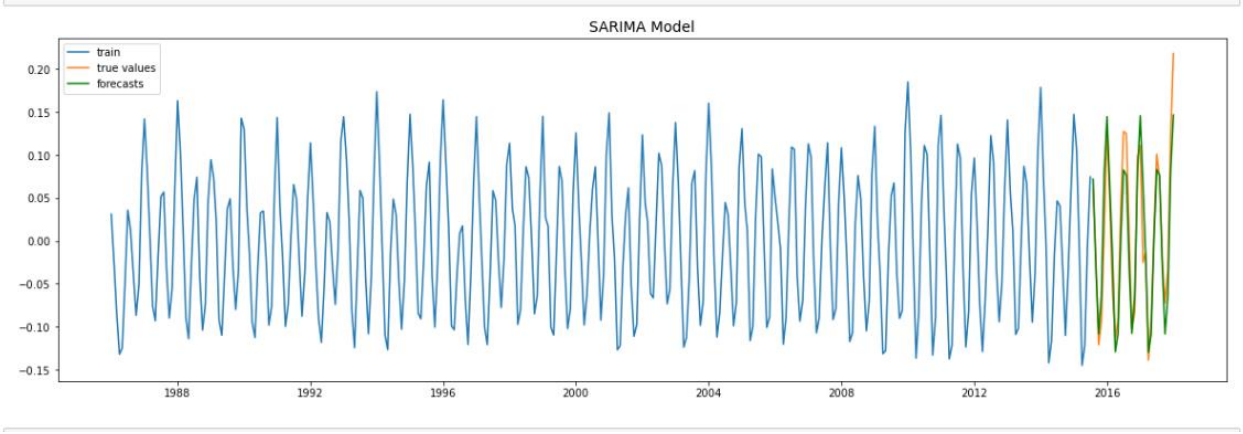


Fig-1: Forecast of 4 years ahead into future

	lower Value	upper Value
2018-02-01	109.485619	119.061210
2018-03-01	99.512463	110.531149
2018-04-01	86.956478	98.456779
2018-05-01	88.247473	99.949544
2018-06-01	98.731406	110.538665
...	...	...
2027-09-01	92.673343	121.080593
2027-10-01	84.514593	113.072616
2027-11-01	87.705332	116.412768
2027-12-01	102.937605	131.793384
2028-01-01	112.513451	141.516661

120 rows × 2 columns

Fig-2: Predicted value

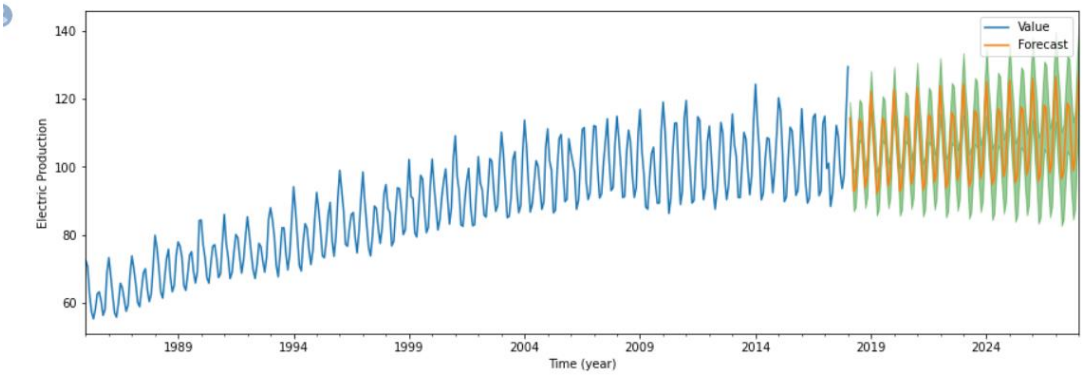


Fig-3: Forecasted value

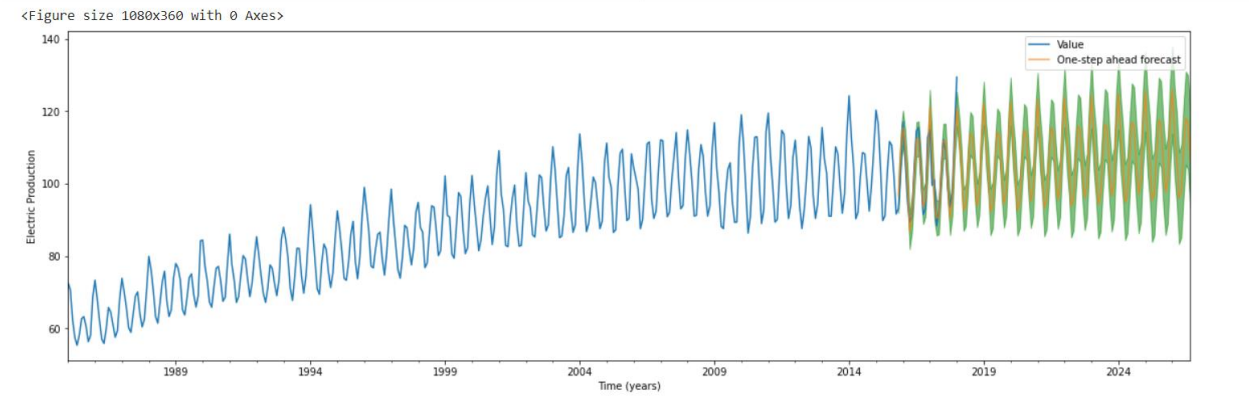
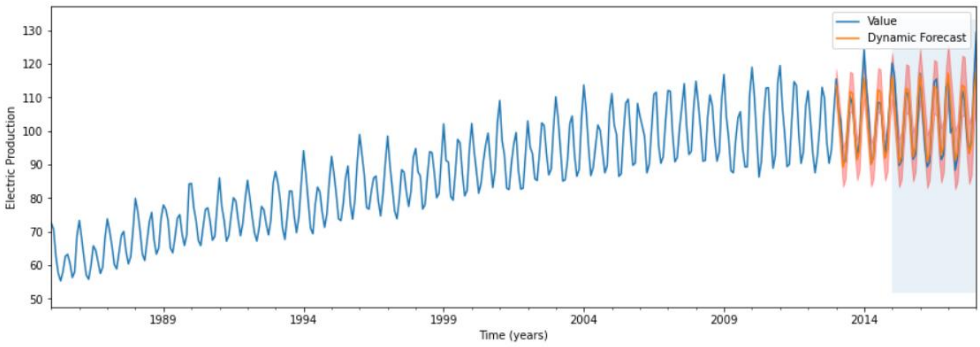


Fig-4: Dynamic forecasted value



COMPARISON OF MSE OF SARIMA, ARIMA, Auto Regression and MOVING AVERAGE

Model	MSE
SARIMA	0.001167
ARIMA	0.004172
Autoregression	0.007904
Moving Average	0.008520

The above table contains MSE of SARIMA, ARIMA, Autoregression, Moving Average in increasing order. We see that the MSE value of SARIMA is the least, hence, SARIMA is the best approach out of all the models

CONCLUSION

From the time series analysis of Electric Production data we can conclude that this project demonstrated the importance of forecasting methods like ARIMA and SARIMA in understanding and forecasting future values of the time series. The results of the analysis showed

that the time series data clearly exhibited seasonal patterns and trends. The ARIMA and SARIMA models were able to capture these patterns and make accurate forecasts.

One of the salient findings of the analysis was the usefulness of considering exogenous variables in the modeling process. When these variables were included, the forecasting accuracy of the models significantly improved. This also highlights the need to consider the influence of external factors on the time series data.

However, the analysis also had limitations. The data used in the analysis was limited to a specific time period, which may have had an effect on the results. Additionally, the choice of model parameters and the utilization of certain assumptions, such as stationarity of the data, could have impacted the results.

Overall, the results of the analysis provided valuable insights into the patterns and trends of the electric production data and demonstrated the potential of forecasting methods like ARIMA and SARIMA for forecasting.

Our future work would include exploring the use of the methods like ARIMA and SARIMA in different contexts or with varied types of data, as well as incorporating additional data sources and variables.