A

Dissertation On (Major Project-II)

“Comparison of AR, MA and ARIMA Process to Predict Price of Copper”

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Aashish Kumar
University Roll No. 2K15/SWT/501

Under the Esteemed Guidance of

Mr. Vinod Kumar
Associate Professor, Department of Computer Science & Engineering

2015-2019(Jan)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, Delhi- 110042

This is to certify that the thesis entitled **"Comparison of AR, MA and ARIMA processes to predict price of Copper"** done by me for the Major project-II for the achievement of **Master of Technology** Degree in **Software Technology** in the **Department of Computer Science & Engineering**, Delhi Technological University, Delhi is an authentic work carried out by me under the guidance of Prof. Vinod Kumar.

Signature:
Student Name
Aashish Kumar
2K15/SWT/501

Above Statement given by Student is Correct.

Project Guide:
Mr. Vinod Kumar
Associate Professor,
Department of Computer Science & Engineering, DTU

ACKNOWLEDGEMENT

I would like to express sincere thanks and respect towards my guide **Mr. Vinod Kumar, Associate Professor, Department of Computer Science & Engineering, Delhi Technological University Delhi.**

I consider myself very fortunate to get the opportunity for work with her and for the guidance I have received from her, while working on this project. Without her support and timely guidance, the completion of the project would have seemed a far. Special thanks for not only providing me necessary project information but also teaching the proper style and techniques of documentation and presentation.

AASHISH KUMAR
M.Tech (Software Technology)
2K15/SWT/501

ABSTRACT

Now we are living in big volume data era, according the era user want a large amount of storage. This was the main problem or all the industries and companies to maintain this large amount of data till 2010. The main purpose was to make a unit or mechanism to solve the above problem in meaning full form. Now we have solve the storage problem with the help of Hadoop and other similar framework on different platforms, and now our storage related problem has been fixed and focus has been shifted to processing data in the form of meaningful form , so data science help to process data and manage in particular order.

Data science is the term in which user can analyses the large amount of data with the help of tools and other mechanism which is present into the market , with the help of those tools and methods user can understand the nature of the data and according the nature it can be analyzed to corresponding use. Data science is used into the many fields to provide to make human like more easily, its helps in mathematics and also help full in statistic information.

Data science used undoubtedly a field which is play an important role in IT fields ,that is taking world by storm but venturing into it because your friends are doing it doesn't make sense. Suppose if you have a series of data and then you want to check or predication the data then its play very important role, data science is help full to understand the nature of data . This task is handled by the Data scientist, it's the major responsibility of the data scientist to perform any task on to the data.

a) Identifying the data-analytics problem to offer the great opportunity for the organization.
b) Determine the correct data set value and variables.
c) Collecting large number of data which can be in the form of structure or it can be in the form of unstructured.
d) It's also help to provide clear data and also check the validation and also guarantee to make data accuracy, uniformly and also provide correctness of the data .

e) Analyzed the data which can be in the form of structure or unstructured format provide or help to check the pattern which is hidden in structure form.
f) Changed the data in a form which is help full to provide solution and opportunities.
g) Also we can find out the data from communication which is held with stakeholder using the tool of visualization and so on.

Above are the tasks which every data scientist have to perform or responsibilities so get the reasonable data for any organization.

Once we gets the data into the real shape means get both type of data like structure and unstructured data, now crucial part is start from here is data analysis, which combines visualization and also from the data sense.

This is most important thing to change the data into a format so that it can change or convert easily so we have there are many source from which we got the data for analysis, these sources are database , flat files present into data base, spreadsheets and data which is not in good or correct format receive from social media.

TABLE OF CONTENTS

CHAPTER 5:

RESULTS AND ANALYSIS

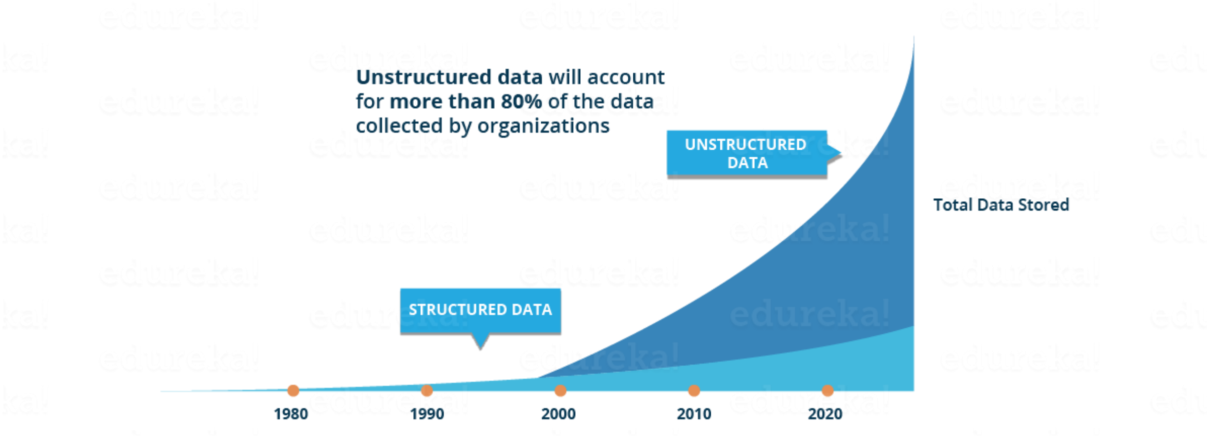
CHAPTER 6:

CONCLUSION

Chapter 1: Introduction

1.1 PROBLEM STATEMENT

Previously, the data which we had was not so much means in previous times we have only small amount of data so to know the nature for that data we can easily or analyzed with the tools after draw the charts and diagrams also in that time data which we had mostly become into the structured form so easily we can draw graph and charts with change the data but this time we have large amount of data which can be in structure or can be unstructured so it's hard to check or predict to previous years. Below is the diagram which is represent the data for this era according the graph data is increasing day by day but not following any patterns or we can say that data which is increasing day by day is unstructured data.



The data which is represent in above graph is collecting from different -2 sources like text files, sensor data, multimedia file data and some financial related data. Tools which are using in previous year to draw the data (in that era mostly data become in the structure form) is not able to process such type of data because it's in unstructured form, so it's hard to draw from basic BI tools. That's why we need some complex and algorithm which will help to understand the data in this era. It's not only the main reason to make data science is to popular but there are many reason also which is mentioned in below so have a look of all the points which make data science so popular and demanding in IT field.

a) In previous era, company didn’t have so much information to understand the test to consumer so now we have many exiting data like customer history and customer past browsing, payment, age, purchase which is done recently, also have income related information now on the base with information available with us, we can guide our system more accurately, also we have those information previously but not vast so it’s also main reason to make it more famous.

b) Let’s discussed with other example to understand the importance of Data Science in human life, suppose we have an auto drive car drive from source to destination and collect many data like sensors data values, radar configuration values, take leaser value and draw a artificial map surrounding around and according this it can make a decision about the speed, turn and drive related information, so it’s also another important area which make data science more efficient and popular.

1.2 RECOMMENDATION SYSTEMS

Recommendation system is an information filtering system which removes unnecessary and redundant information and predicts the values for given source.

Consider a manufacturing plant that has been in this market to manufacture the product form last five to ten years, and it has been launched some latest model that will be manufactured in the last few months, now that plant has two way to manufacture the latest laptop models, first is company can check or identify the correct or reasonable market for that product, or it can check the market requirement for next five year based on the previous product data which is available with the manufacturing plant.

So for predication the data we have above two method for production for laptop or any other thing which is beneficial to the consumer, first method is so tough as compare to the second one because every time consumer taste become changes, so any company and industries second will be the best way to predict the data for future references,

Many Manufacture Companies check all the parameters related to the economics like unemployment rate, GDP and so on. Those parameters play and important role for any manufacture company so above parameter also consider so forecast the future values.

There are many scenarios for any company which contain forecasting variables depends upon the time or we can say that it can very time to time so for those company forecasting values for such parameters can be in days, month and years or even it can be second and millisecond, or simply so to represent the time line it can be any unit. Then for a given series if we will predict the future values then for this we will use the time series analysis.

1.3 THESIS MOTIVATION AND GOAL

Recommender systems mostly use historical data to predict the future variable or values. Such systems have widespread usage in e-commerce, stock, and production for any marital or thing.

For example suppose any mobile company want to launch any model into the market so first concern for that company that how many model have to build up so that most of model can sell out into the market so for this in my recommendation system we find out the previous year’s data and check the data form weather it’s in structure form or unstructured form, after the form of previous data into the structure format and predict the future values for the given previous year variables.

Throughout this thesis, we have tried to find out the correct future prediction values, which will help to get the correct future value for prediction.

1.4 THESIS STRUCTURE

This thesis is classified into six different chapters.

Chapter 1 defines the problem statement for the thesis, in which I have described what was the problems which we are facing with the data which we have found form with the data which we got from the social media or from other resource.

Chapter 2 describes the related work done in the areas of deep learning for how to data change in to the structure format and which model is appropriate to change the data into a from which is supported by our models. This thesis deals with many models like AR, MA, ARMA and ARIMA which we will help to make data series as stationary form.

Chapter 3 explains the research topics used in detail. In our research details, we explain the terms, which are being used in the thesis like Auto Regression, Moving Average, Auto Regression Moving Average, Auto Regression Integrated Moving Average, Cosine Similarity, Statistical Features like Mean/Median, Standard Deviation, Histogram values, Root Mean Squared values etc.

Chapter 4 is explaining the solution architecture of the different 2 model which is used for predict the price of copper and explains about the algorithm used to make the date set as stationary so that we can perform the operations to predict the values for future use. It deals with the extraction of use data from the given set of data, where we can check the behavior of the data which can impact on to the business. With Support development in the area of text mining or large data, we have worked on a many models so check which model provide the correct result and which model giving less accuracy . In this chapter we have arrange the data into the format so that we can perform many operation so that in this our first task is to make the data set stationary so that we can provide this stationary series to our model to get value. So in this we describe the method or diagram to check the stationary factor weather its stationary if it’s not stationary then we perform many operation to make it stationary.

Chapter 5 illustrates the step by step results of the complete approach starting from processing of the data to the result comparison of the predicted values with the original values. In this step we have used a clear approach and according this we get the series front the data and make diagram check the stationary factor and all the results are shown in the form of graphs and plots for better visualization. Finally the implications of the results are discussed in this chapter.

Chapter 6 is the conclusion of the thesis. It describes the benefits of the Box-Jenkins approach for predictive analysis of the time-series process, and also described the result with the graph also discussed about the accuracy of the approach with different 2 model used in this predication. Also the future work which can be done on the current work.

Chapter 2: Related Work

[1] Airline Passenger forecasting using the neural network and Box-Jenkine. In this demand forecasting for the remaining or available seats into the airlines is important to maximize the expected revenues by setting the appropriate fare levels for those seats.

[2] Time series forecasting using a hybrid ARIMA and neural network model, ARIMA is one of the popular linear models in time series forecasting in past three decades, in this ARIMA model and ANNs are often compared with mixed conclusions in terms of the superiority in forecasting performance.

[3] Statistical signals processing, detection, estimation, and time series analysis, in this analysis the probability of signal detection and check the future possibility with the time series to get future predictions values.

[4] ARIMA implementation to predict the amount of antiseptic medicine usage in veterinary hospitals, in this a healthcare institute such as pharmacies or hospitals must ensure the availability of medicine for the patients.

[5] A hybrid ARIMA and support vector machines model in stock price forecasting, in this ARIMA model cannot easily capture the nonlinear patters, Support vector machines or a novel technique has been successfully.

[6] Prediction of rupiah against the US dollar by using ARIMA, In this currency exchange rate is needed in the business word for example, investment and profit assessment, Prediction of rupiah rate is done to get the price of rupiah against US dollar in the future to be used as consideration in decision making about when will be the correct time to investment.

[7] Short term traffic flow prediction using a methodology based on ARIMA. Accurate short term traffic flow forecasting is fundamental to both theoretical and empirical aspects of intelligent transportation system deployment, in order to play ARIMA model with good linear fitting ability and artificial neural network model with strong non-linear relation mapping ability, this study aimed to develop a simple and effective hybrid model for forecasting traffic volume with that time slots.

[8] Forecasting method of aero-material consumption rate based on seasonal ARIMA model. It is indispensable to scientifically predict the consumption of aero-material and to make scientific decisions on aviation equipment maintenance resources and make full use of existing resources to improve maintenance capability. In the process of aviation equipment maintenance and support, the consumption of aero-material tends to show a seasonal change. This paper proposes a seasonal ARIMA (Autoregressive Integrated Moving Average) model to solve the problem of aero-material consumption rate forecasting

[10] Forecasting of raw material needed for plastic products based in income data using ARIMA method. Forecasting is a process of predicting something future by doing calculations from previous data. In this case the authors will forecast the sale of plastic production by using ARIMA Box-Jenkins method for 2015 forecasting. The data used is the sales data of plastic factory production in Bandung from 2012 to 2014.

[11] Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network. Stock price index is a barometer of the national economy, which often shows strong nonlinearity because of various factors. It is necessary to use nonlinear models to improve the accuracy of prediction. We predict Shanghai Securities Composition stock index with ARIMA-BP neural network method, and compare the accuracy with the result of single ARIMA model and BP neural network method. We find that the prediction accuracy of ARIMA-BP neural network is better than the BP neural network, BP neural network is better than linear model ARIMA, which confirms the change of stock price index is nonlinear.

Chapter 3: Research Background

3.1 Time Series Process

Time Series process is a process which has data values differs with each other over a time of periods or it's a sequence of observation of value of data varying with the time. And time period in which data is vary, it can be fixed or can be vary with some value but other manipulation the time, we will put equal time of intervals in which data varying in this process. Data like the number of flight booked for an airline in a month or number of loans applied to a particular bank in a month, all examples like these come under the category of the Time Series process. If these companies are successfully able to predict the future values of number of flight booking or number of loans that will be applied in the upcoming time interval, this predication if done successfully can be very helpful for the companies in allocating funds and resources and efficiently meet the requirements.

3.2 Time Series Analysis

Time Series process undergo Time Series analysis so that using the previous data future prediction can be made and useful patterns can be drawn from the historical data that could be beneficial in future dealings. Time Series involves three major steps; Descriptive analysis, modelling and forecasting the future values. First step that is the descriptive analysis is all about understanding the properties of the Time Series process in hand. It is about looking for trends, seasonality and the behavior of the series. Based on the first step, the second step of modelling is performed. Identification of the correct technique to be used to prepare the model is finalized. Once the model is fully prepared using the training data, the next steps comes into the picture that is forecasting the values using the prepared model in second step. Using the test data set forecasting power or accuracy of the model is tested. If the model does not meet the required threshold of accuracy the model is rebuilt. Hence it is an iterative process until the desired accuracy is achieved.

Basically it divided mainly in 3 phases of time series analysis:

a) Descriptive analysis

b) Modeling

c) Forecasting

So in Descriptive analysis we have to find out the type of the data , tried to find out what is the sequence of data weather it is in structure or unstructured , or check if there is any sessional or trend dependencies or not, and in Modelling step . Once we identified the patterned then we will forecast the future values.

Basically in time series we will use the above three step to forecast the data for future values. In above phase second phase play a important role to cast the future values, so in to the market there is many tools present to modeling the time series analysis but here I have used Box Jenkines approach (ARIMA) , it’s very use full and frequently used in time series

3.3 Box-Jenkins Approach

One of the most popular techniques for Time Series analysis is Box-Jenkins approach. Box-Jenkins approach famous the other techniques also, but mainly its auto-regressive integrated moving average (ARIMA) technique of modelling a time series. Box-Jenkins approach did significant advancements in the ARIMA method by simplifying the application of the method. There are many fields where this approached has proved to be very useful in accurate prediction of the time series process. Many work has already been done using this approach and has shown very positive results. One needs to have good understanding of AR, MA, ARMA and ARIMA processes for implementation of Box-Jenkins approach to do predictive analysis of Time Series processes.

Time series models are known as ARIMA (auto regression integrated moving average) so ARIMA model is divided in to the three parts:

a) AR (Auto-Regression)
b) I (Integration)
c) MA (Moving Average

3.3.1 AR (AUTO-Regression)

In to the AR process where the current values which is using for any operation depends upon the previous value which is already used, and its denoted by AR(k), where k is the order of auto-regression process. P factor also denote the relation between the current value and the previous value so its determine how they are depend with each other’s

Suppose we have one time series like Xt , Xt-1, Xt-2, Xt-3, X-4, Xt-5 , Xt-6, . . . Xt-k, which is a time series and if the series is an AR(1) series , then the value of Y at the given time t will be:

$$Xt = b1*Xt-1 + et$$

Where b1 denote the quantified impact factor, and et denote random error at time t. et is also known as white noise. For an AR(2) series,

$$Xt = b1*Xt-1 + b2*Xt-2 + et$$

Where b1 denote the quantified impact factor, and b2 denote quantified impact Xt-2 on Xt .Likewise, for a series to be AR(P), Y will depend on the previous Y values until the time point t-p.

$$Xt = b1*Xt-1 + b2*Xt-2 + b3*Xt-3 + ...+ ak*Xt-k + et$$

b1, b2, b3, ... ap are the quantifies impact of Xt-1 ,Xt-2, . . . Xt-p on Xt

3.3.2 MA (Moving Average)

MA process is the process in which the previous values of the time series process do not have any effect on the current values of the process, rather current values depend upon the error or noise of the previous values.MA (p) is the symbol used to represent the MA processes. P denotes the number of previous values on which the current value depends.

For above given series if it’s MA (1), then at the time t to t-1 deviation will be.

$$Xt - g = et + b1*et-1$$

• b1 denote quantified impact of et-1.

• g its mean.

• Xt - m it t time deviation.

If series is MA(2), then at the time t-1 to t-2 deviation will be.

$$Yt - m = et + b1*et-1+ b2*et-2$$

In the previous equation:

• b1 and b2 are denote quantified impact of et-1.

• m its series mean.

• Xt -m at t time deviation.

Similarly, for an MA(q) series, at the time t, t-1, t-2, t-k deviation will be

$$Xt -m = et + b1*et-1+ b2*et-2. . . ..+ bq*et-q$$

So now to make this model to effective we will integrated both the process AR + MA so now we will get another new process, and this process is known as ARMA.

3.3.3 ARMA (Auto Regression Moving Average)

ARMA process depicting behavior of both AR process and MA process, hence called ARMA process. So there is both long term effect on current values of previous values as well as short term effect of noise of previous values. ARMA (p, q) denotes an ARMA process where p depicts order for AR process and q denotes order for MA process.

Consider if Xt , Xt-1 , Xt-2 , Xt-2 , Xt-4 , Xt-5 , Xt-6, ... , Xt-h are the values for the time series data set and the errors at time t, t-1, t-2, ... t-k are et, et-1, et-2, . . . , et-h respectively, then an ARMA(1,1) as follows:

$$Xt = a1*Xt-1 + et + b1*et-1$$

• b1 denote quantified impact of et-1.

• g its mean.

• Xt - m it t time deviation

An ARMA(2,1) series as follows:

$$Yt = a1*Yt-1 + a2*Yt-2 + et + b1*et-1$$

3.4 Working of Box-Jenkins Approach

In this approach to forecast firstly we have to make sure the series in which we are working is stationary or not stationary, then we need to check the current type of process weather its AR process, MA process or ARMA process. For its value prediction we will follow below two steps:

a) Testing weather the Time Series is stationary: firstly check weather given series is stationary or not if not stationary then we have to make it stationary.to check weather its stationary we use DF(Dickey Fuller) Test, After perform this test getting the value of p and check it this value of p greater than 5% then series which we have got after the operation is not stationary if value of p less than 5% then its stationary.

b) Second we have to check order of the process, and for this we are using Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF).

3.4.1 ACF (Auto Correlation Function)

ACF (Auto correlation function) denote the correction for two values with in some time interval suppose we have two variable with some time difference such as t and t-1 denoted as Xt and Xt-1, if we will try to find out correlation between X and the previous value which is already used and them the correlation will be the at different lags

• Auto correlation function with lag zero denoted by ACF(0): its denote at lag0(p0) Correlation will be = Xt and Xt =1

• Auto correlation function with lag one denoted by ACF(1): lag1(p1) correlation will be equal to the relation with in Xt and Xt-1

• Auto correlation function with lag two denoted by ACF(2): lag2(p2) correlation will be equal to the relation with in Xt and Xt-2

• Auto correlation function with lag three denoted by ACF(3): lag3(p3) correlation will be equal to the relation with in Xt and Xt-3

3.4.2 PACF (Partial Auto Correlation Function)

PACF (Partial auto correlation function) its similar to the ACF but having some difference in ACF we get the correlation of two variable with the change in lags but in PACF we have to remove effect to other variable to the current values .

In this PACF we get the the current value after taking lags with the previous values

• Partial Auto correlation function with lag 0 denoted by PACF (0) : its denote at lag0(p0) Correlation   will be = Xt and Xt =1

• Partial Auto correlation function with lag 1 denoted by PACF (1): Calculated PACF at the time when lag is equal to one and its denoated by lag (p1) = it will the coefficient of previous value of Xt-1 when Xt-1 goes to Xt.

- Partial Auto correlation function with lag 2 denoted by PACF (2): Partial Auto correlation function with lag 2 denoted by Lag (p2) = it will the coefficient of previous value of Xt-2 when Xt-2 goes to Xt.

- Partial Auto correlation function with lag 3 denoted by PACF (3): Partial Auto correlation function with lag 3 denoted by Lag (p3) = it will the coefficient of previous value of Xt-3 when Xt-3 goes to Xt.

3.4.3 IACF (Inverse-Auto-Correlation Function)

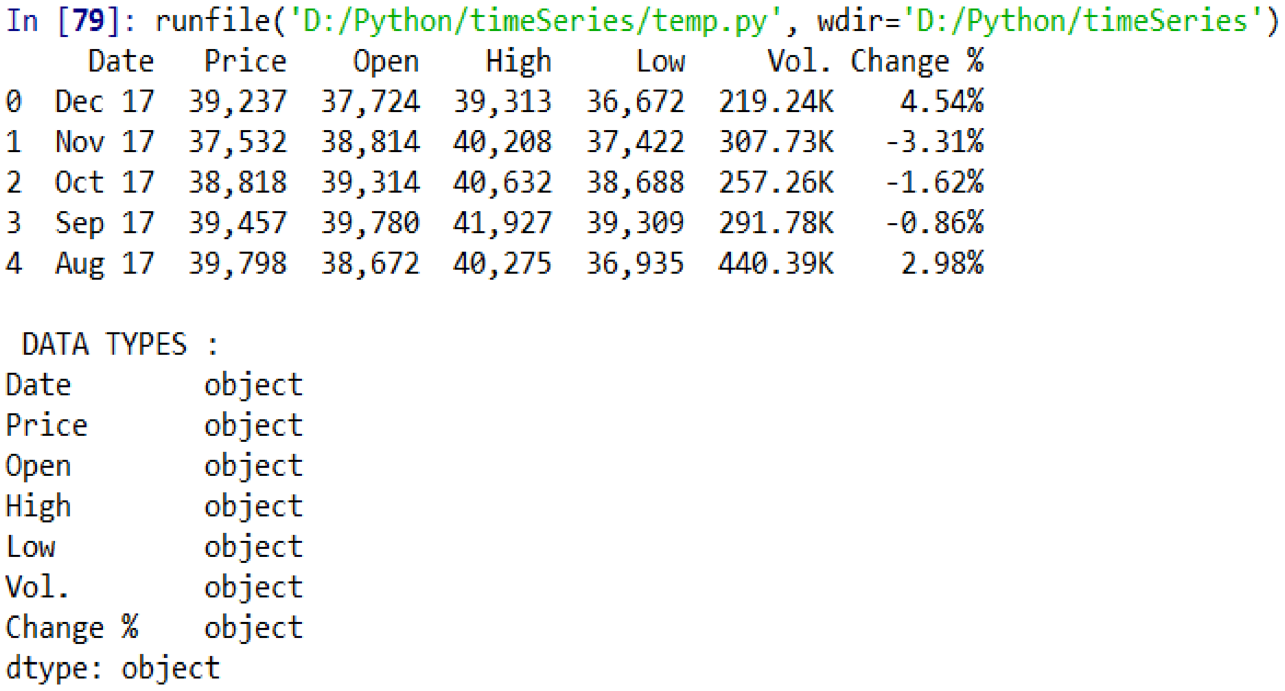
The inverse auto correlation is different from the ACF and PACF ,in this next value doesn't depend upon the previous value , but its related to error in to the series , in this we have to change the functionality or role and the series which we are getting for processing , in this if will draw the auto correlation function for any value for particular order , consider we are drawing (2,1) then it will be same draw or will be look like to Inverse auto correlation function with the reverse order of ARMA , then it will be equal (1,2). Also this type of Graph is helpful to check, is there any trend is present in to given data set.

Chapter 4: Proposed Approach

In first step we have list of copper data which include several years' data with corresponding date and price value, so firstly we will print the data:

4.1 Collection or Data Gathering Process

In first step we have list of copper data which include several years' data with corresponding date and price value, so firstly we will print the data:



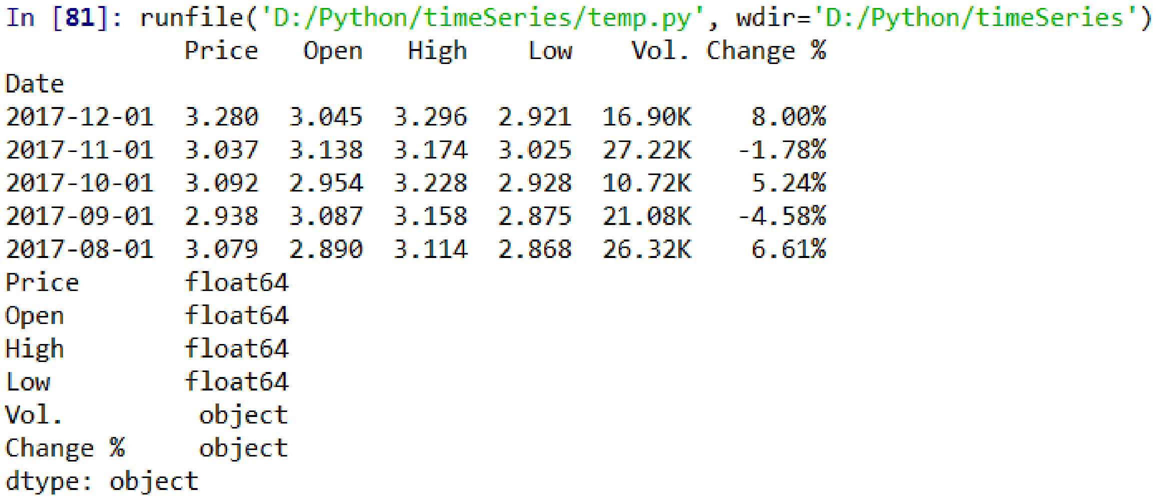
```
In [79]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
     Date   Price    Open    High     Low     Vol. Change %
0  Dec 17  39,237  37,724  39,313  36,672  219.24K    4.54%
1  Nov 17  37,532  38,814  40,208  37,422  307.73K   -3.31%
2  Oct 17  38,818  39,314  40,632  38,688  257.26K   -1.62%
3  Sep 17  39,457  39,780  41,927  39,309  291.78K   -0.86%
4  Aug 17  39,798  38,672  40,275  36,935  440.39K    2.98%

 DATA TYPES :
Date        object
Price       object
Open        object
High        object
Low         object
Vol.        object
Change %    object
dtype: object
```

Fig 6: System Architecture of Friend Recommendation

In this simply are printing the data from the xml value, this data having columns such as Date, Price , Open ,High, Low, Volume, and Change and those are the data types which are using in in our code. So this time data having their own index values which is start from 0, 1,

4.2 Organize the Data in supported format

In second step we will organize the data in the form which is supported by the python, in above diagram the index number is default occurring so now we use the date as an index and change the date in to proper format supported by Python. Also in previous diagram all the data type are objects type so have to change data type to the corresponding types so that we can easily manipulate those value.

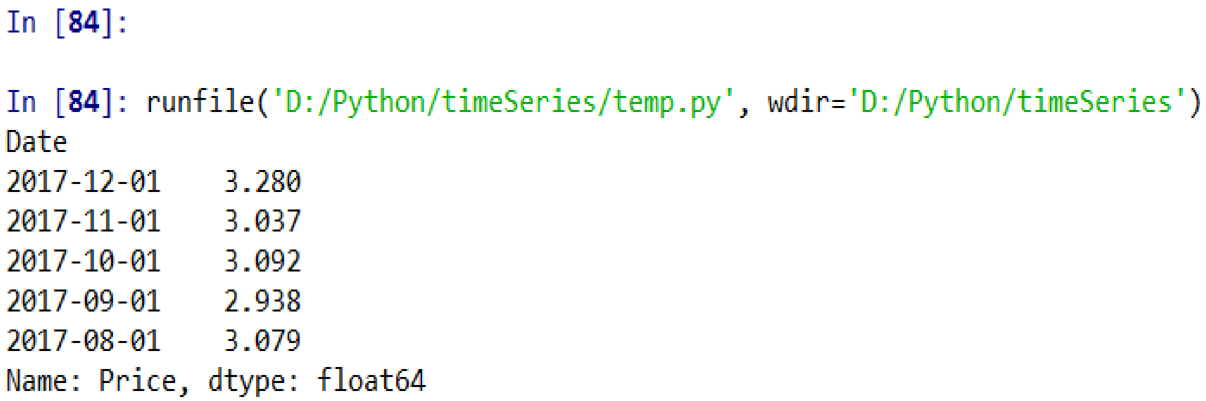


```
In [81]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
            Price   Open   High    Low    Vol. Change %
Date
2017-12-01  3.280  3.045  3.296  2.921  16.90K    8.00%
2017-11-01  3.037  3.138  3.174  3.025  27.22K   -1.78%
2017-10-01  3.092  2.954  3.228  2.928  10.72K    5.24%
2017-09-01  2.938  3.087  3.158  2.875  21.08K   -4.58%
2017-08-01  3.079  2.890  3.114  2.868  26.32K    6.61%
Price       float64
Open        float64
High        float64
Low         float64
Vol.         object
Change %     object
dtype: object
```

So now we have the value having correct and supported format.

4.3 Manipulation Row Data

Now this Step we will extract the value which we want to use. This data having many rows but currently we extract those row in which we want to use in our code, so here we will use date and price so the resultant data are:



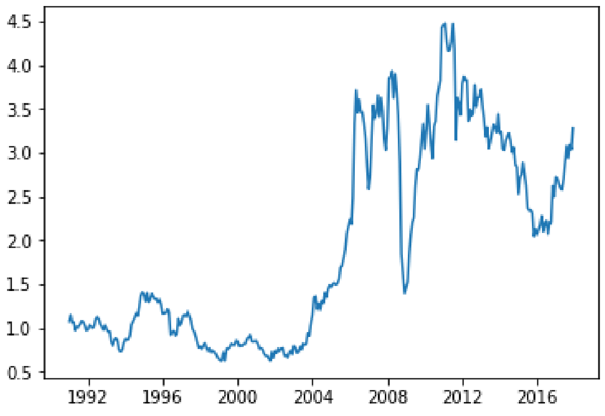
```
In [84]:

In [84]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
Date
2017-12-01    3.280
2017-11-01    3.037
2017-10-01    3.092
2017-09-01    2.938
2017-08-01    3.079
Name: Price, dtype: float64
```

So above table having resultant data in which we manipulate the data.

4.4 Draw Diagram

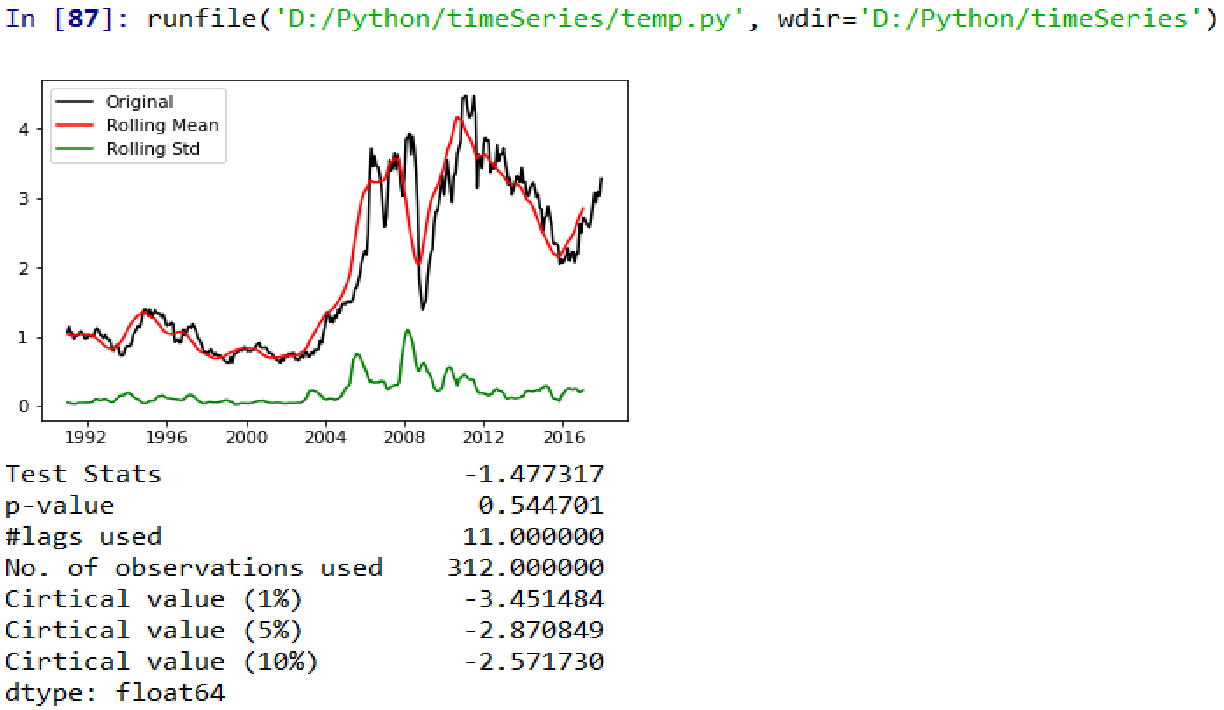
In this Step we will check that series which we have received from above method is stationary or not, so before check stationary we will draw a graph with year and rate change

```
In [86]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```



4.5 Check Stationary factor

Now in this step we will check whether we have stationary data series or not for this we will the stationary test on the given data and got below values:

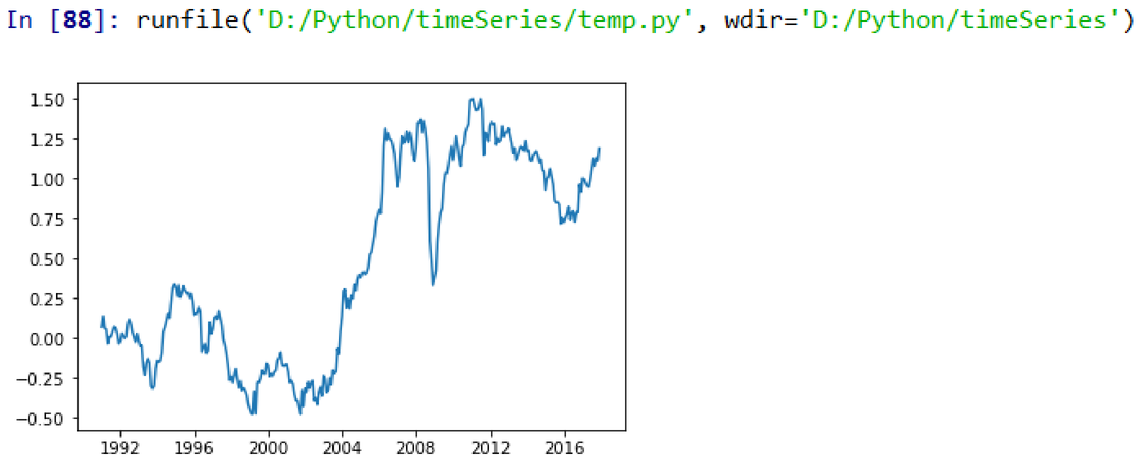


```
In [87]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

```
Test Stats                     -1.477317
p-value                         0.544701
#lags used                     11.000000
No. of observations used      312.000000
Cirtical value (1%)            -3.451484
Cirtical value (5%)            -2.870849
Cirtical value (10%)           -2.571730
dtype: float64
```

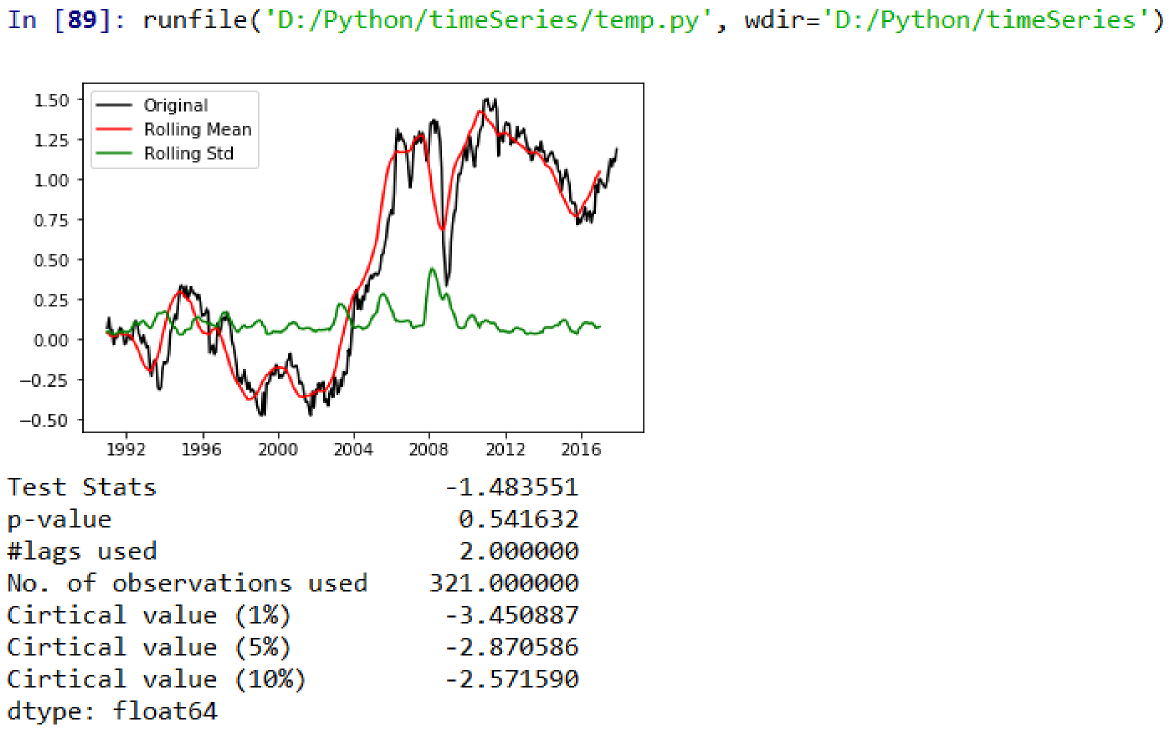
In above data we will check the value of p if value on p is less than 5% then the series is stationary and greater than 5% then it's not stationary but in our data we got the value of p is 0.544701, it means value of p is approx. 54%, so current series is not stationary so firstly we make that series stationary.

So to make series stationary so have to take log function , it means we will try to take a small part of the series which is stationary, so after taking we will plot the graph of the data which we will get after taking log function:



```
In [88]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

Now again we will perform stationary test on the data which we get after use the log function and check value of p , if value of p less than 5% then the resultant series is stationary otherwise not stationary so after perform the stationary test we will get the below Graph:

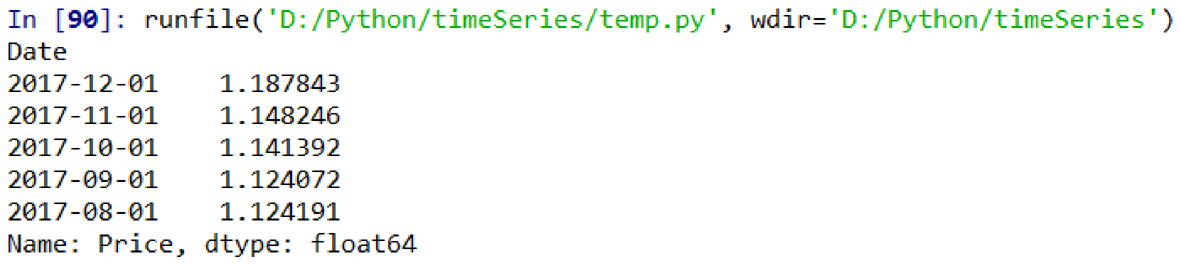


```
In [89]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

```
Test Stats                       -1.483551
p-value                           0.541632
#lags used                        2.000000
No. of observations used        321.000000
Cirtical value (1%)              -3.450887
Cirtical value (5%)              -2.870586
Cirtical value (10%)             -2.571590
dtype: float64
```

So now the value of p is .541632 and it’s near to 54%, it means series is not stationary, so our main task is to make series stationary.

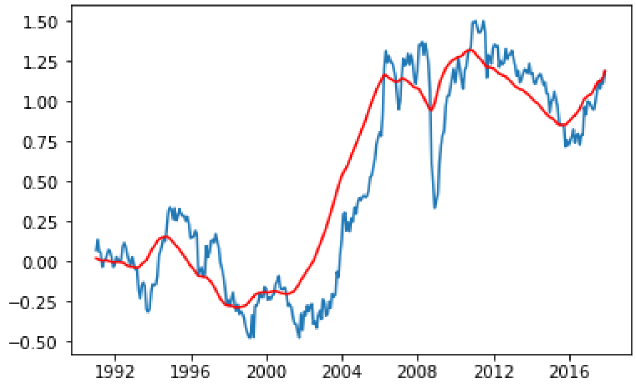
So now we select a small part of series which is stationary so for this we will get the exponential weight of the series, so after getting exponential weight of the series, resulting data are below:



```
In [90]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
Date
2017-12-01    1.187843
2017-11-01    1.148246
2017-10-01    1.141392
2017-09-01    1.124072
2017-08-01    1.124191
Name: Price, dtype: float64
```

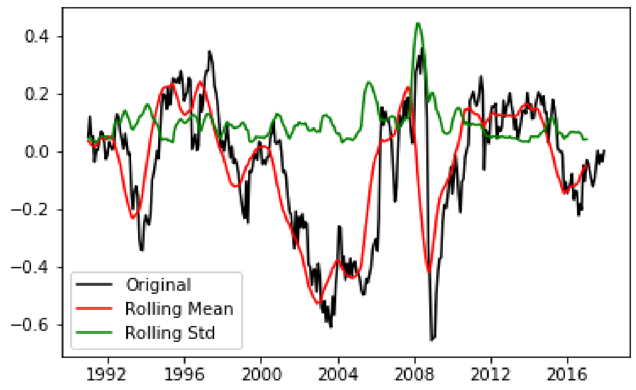
So now we will check the difference between original data and exponential weight from the below graph:

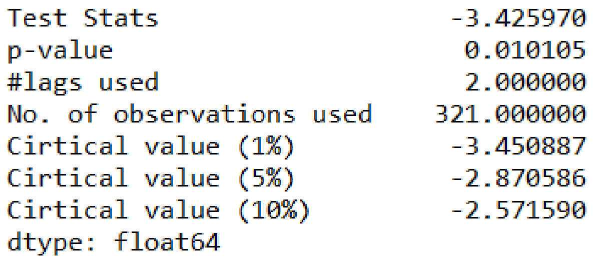
```
In [92]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```



Both the graphs are different and means both the value means original in which we are working and value get after exponential function having a lot diff, so now we will perform stationary test on the data which we got after perform exponential function.

```
In [93]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```



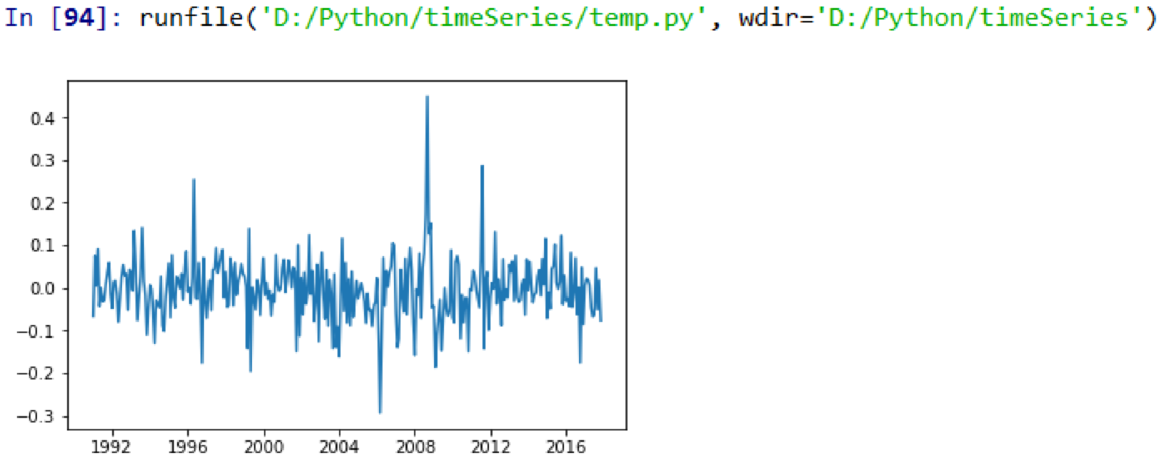


```
Test Stats                      -3.425970
p-value                          0.010105
#lags used                       2.000000
No. of observations used       321.000000
Cirtical value (1%)             -3.450887
Cirtical value (5%)             -2.870586
Cirtical value (10%)            -2.571590
dtype: float64
```

Here we got the value of p is .010105 ,its means the value of p is approx. 1% , and its less than 5% now we can say that series which we found after exponential is stationary.

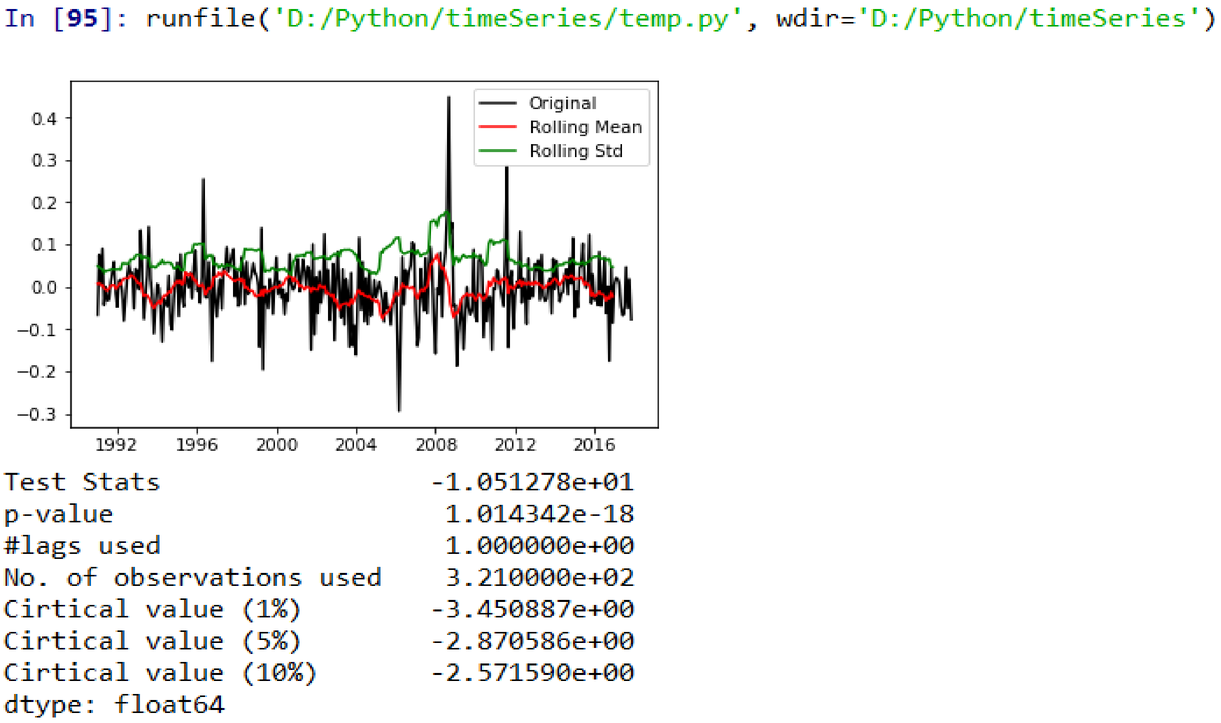
4.6 Check Trend and Seasonality Effect

But there are many factor which we have to consider also like trend and seasonality, so in our prediction we have to consider those value also, so for this use technique differencing which tackle both treads and seasonality, so after differencing we will get the below graph:



```
In [94]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

So after getting the differencing, now we will check whether the series which we got after differencing is stationary or not so we will perform text on differencing series:



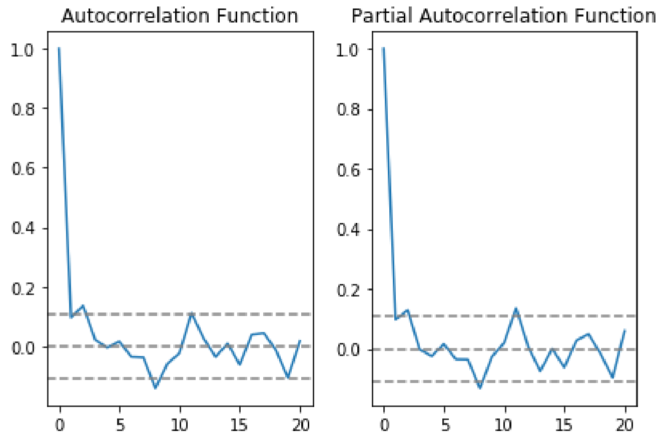
```
In [95]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```

```
Test Stats                       -1.051278e+01
p-value                           1.014342e-18
#lags used                        1.000000e+00
No. of observations used          3.210000e+02
Cirtical value (1%)              -3.450887e+00
Cirtical value (5%)              -2.870586e+00
Cirtical value (10%)             -2.571590e+00
dtype: float64
```

Here we will check the value of p and in above diagram value of p is 1.014342e-18, and this value is in negative and it better than the previous series so for our manipulation we will consider differencing series as resultant series.

Now we have more stable stationary series which we achieved from differencing methods. So now we will check the order of series and will we draw the ACF and PCF graph to achieve the order of process:

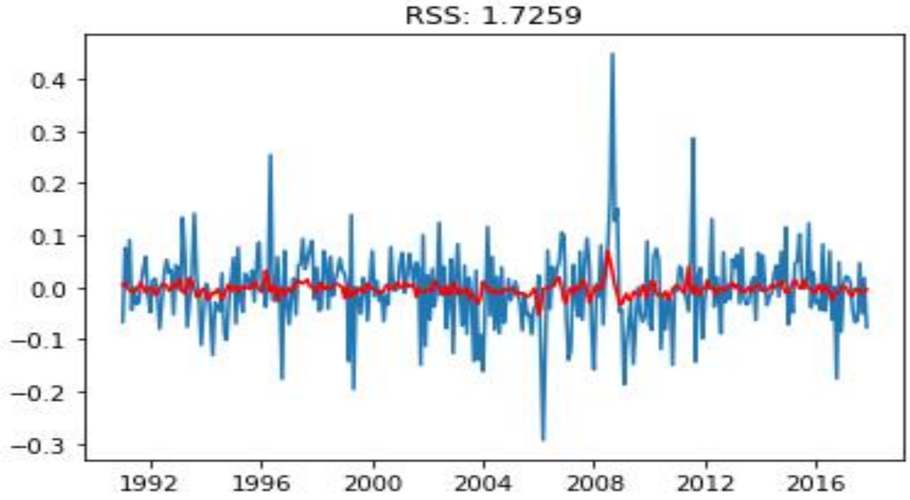
```
In [97]: runfile('D:/Python/timeSeries/temp.py', wdir='D:/Python/timeSeries')
```



So now we have to check which process will be more suitable for our series and we will draw all the three process like AR process, MA process and ARMA process and check which process having less RSS value.

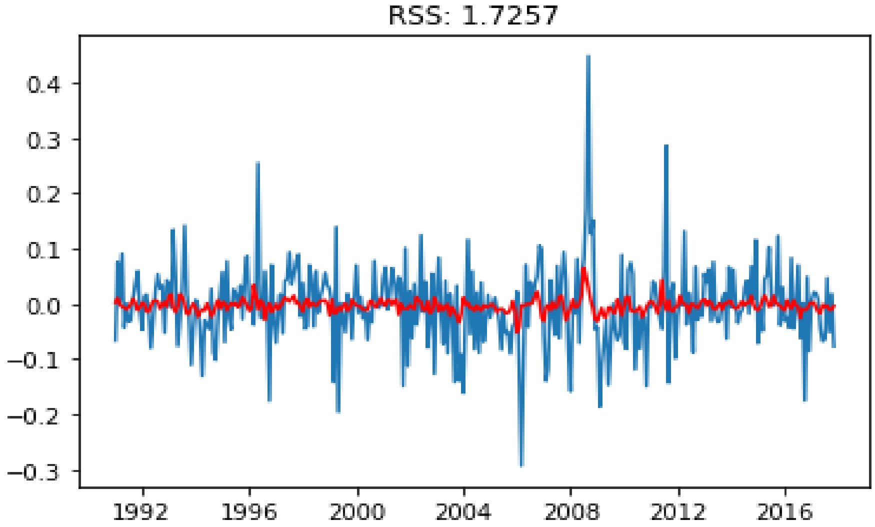
4.7 AP Process Diagram

So in AR, the next value depend upon the previous value so AP process diagram for the above series:



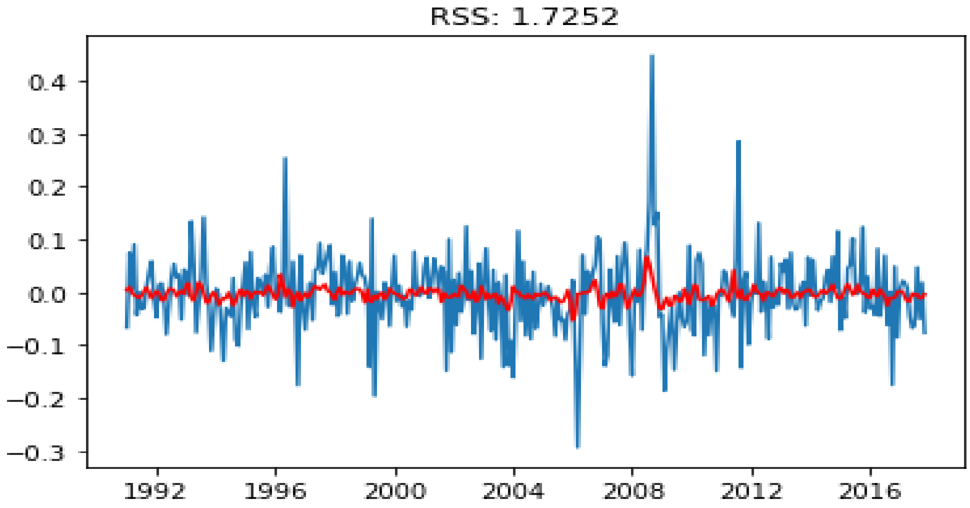
4.8 MA Process Diagram

so now we will draw the MA process graph in which previous value depend upon the error factor so graph so MA process is:



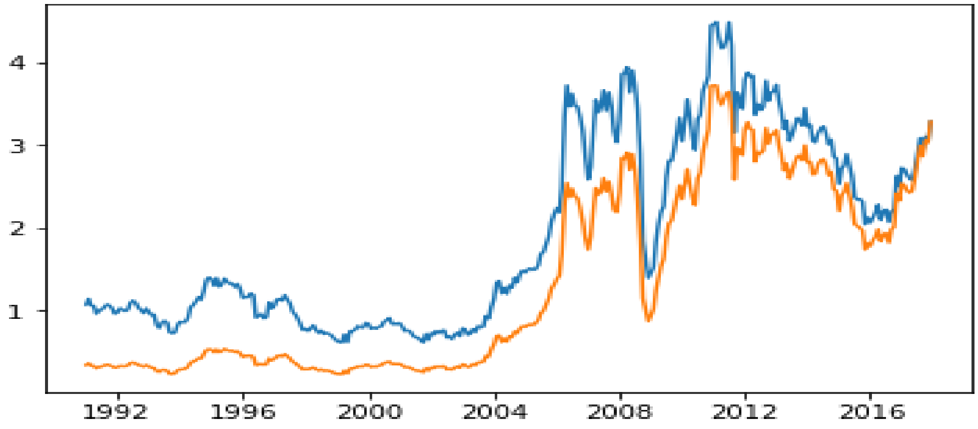
4.9 ARMA Process Diagram

So in we will draw the graph of last process which ARMA process, in this process we marge both the process and will add both the process advantages. So resultant Graph of ARMA process is:



Now we have all the three and process graph along with RSS factor value so in AR factor value is 1.7259 and in MA process RSS factor value is 1.7257 and in ARMA process factor value is 1.7252 so in all the above process we will select that process which having less RSS factor value, so according our data we have ARMA process having less RSS factor value them we will consider ARMA process for the data set which I have taken.

Now we have all the value means stable stationary time series and order number from ACF and PACF and process which we will use for our data set then we will draw our prediction below:



So in above diagram there are two lines are one is blue line which is draw from actual data set and other one is yellow which is draw from the prediction data set, So from above diagram we can say that our prediction value is not same as original (having small difference) but check the trend on graph which is similar to the original value.

Seasonal dependency is another factor which affect the result so if we are going to analysis for series of predict the future variable value then it will also affect our result also so in this we have to consider it during computation the result. Consider with one example, IN India there is a festival call Diwali, in this festival people purchase the steel material so if some manufacturing plant developing any steel material then those manufacture plant want to predict that how much production will be enough for this year and for that prediction that Manufacture Company have to consider this seasonality in their code and produce the result and that result will be good as previous result,

ACC (Autocorrelation correlogram): for a give series if we want ot find out the season pattern then it will be represented in to the form of correlograms. And for this if we want to show the value of seasonality then we show it by graphically with the help of ACF. We are using ACF so show the seasonality because its show that seasonality with some specific lag or we can say that there will be an lower and upper bound for lags. So if we want to show the difference between to the lags and for this we have to draw correlograms diagram. For this we are using only auto correlation function because its more easily find out the lags and easy to understand the lags difference which is play an important role for to predict seasonality factor.

Examining correlograms: For understanding the correlograms first we should understand the functionality of ACF because its major one, also there is one thing should be in mind the consecutive lags values in ACF show dependency with each other. To understand this concept let us an example , suppose we have three element and first element having some value which is depend upon to the second element and second element having some value which is depend upon to the third element values so now we can say that first element value depend upon the third element value and vice versa. So form the above example we can say that if we want to remove that dependencies between the element and we can remove it by ACF (1).

Chapter 5: Results and Analysis

5.1 Data Collection and Preprocessing Details

We can apply ML to all the times series data value which can be in any form either it can be structure or it can be unstructured or value which we are got from the data source. These value can be easily change into the structure form or we can divide the whole series with in time interval means it's maintain the data with in time interval. It's a problem with machine that it's required data in to the desired format for process.

In this, you will divided the classification in 8 examples to get the time series values so that it can work as an input for the machine. These are below define,

a) 4 univariate time series datasets values.
b) 3 multivariate time series datasets values.
c) During browsing we can download the all the string or keyword which are using to search on internet or any other resources.

5.1.1 Uni-variate Time Series value data sets

In this we have include that series which is distinguish with single variable or we can say that this type of series having only one type of parameter , this type data series is known as Uni-variant Time series data value . These datasets having below properties for classification are mentioned:

a) This type of series is simple in structure and clear sequence of data and user easily predict the value.
b) We can easily figure out the data and draw the diagram for clear result.
c) After drawing the value by a diagrams we can easily check value from the prediction value.

d) After drawing those value which is figure out by only one variable, from the diagram we can easily check the traditional trend and find out the new method which we can imply on it.

Data which we got after browsing during search any value on internet that data set in both the format which can be structure and unstructured format. Specifically, also in in data science we have some times series analysis library is already there which is developed by the developer to make it as open source.

Mentioned the some date set values that we are getting easily for the website search data with the sales data, Meteorology, Physics and Demography.

a) Shampoo Sales Dataset values
b) Monthly Sunspot Dataset values
c) Daily Female Births Dataset values

5.1.2 Multi-variate Time Series value Data sets

It's basically more difficult as comparison with univariate date set and those are the easily used variables for ML libraries which process those type data and provide the result according to the process there are many sources from which we can get those type data also in above data set we used only one variable but in this it's can be more than two variable so in short we can say that this type of data set is more complex than the previous one. We have in our libraries 63 values of data sets we can download and can update means read and write the values.

Below I have mentioned three data set which we can used for our work for any purpose

a) EEG Eye State Values
b) Occupancy Detection Values
c) Ozone Level Detection Values

5.2 Identification Phase

5.2.1 Parameter Estimation Process

Once the time series is made stationary, comes the important step to identify the type of time series process. This step is important as correct identification of process determines the accuracy of the model prepared for prediction of values of the time series process. One need to identify out of AR, MA, and ARMA process which category the times series process belongs. Once the process has been successfully identified, next is to predict the order the process finalized. Also there are some parameter which we need to check like trend and seasonality related parameter value. So before prediction any value we have to check whether we have included this factor or not.

Identification of the process type cannot be done by visually analyzing the graph of time series process. Certain metric are required to correctly identify the process type. PACF and ACF functions and plots are used to identify to which process the time series truly belong.

For identification of the AR process ACF function and its plot is used. The rule to be followed is to check the ACF plot, if the plot dies down to zero or shows a tendency of reducing to zero then we can safely say that the time series process belongs to AR process. Now to identify the order of AR process, PACF plot need to be checked. The point where the PACF plot cuts off indicated the order of the AR process.

For identification of the MA process PACF function and its plot is used. The rule to be followed is to check the PACF plot, if the plot dies down to zero or shows a tendency of reducing to zero then we can safely say that the time series process belongs to MA process. Sometimes PACF plot does not clearly show the trends, hence IACF plot is used in that case. IACF plot is nothing but just inverse of ACF plot. Now to identify the order of MA process, ACF plot need to be checked. The point where the ACF plot cuts off indicated the order of the MA process.

5.2.2 Seasonal Related Models

This is the factor which effect the our result more, why its effecting more so for this let consider an example, suppose there are some product which is sell out most in some particular time so id some product company manufacture that product and before producing anything, that company will decide how much production have to do to earn more profit so that all product will be sell out. So for this company take some previous years data and according the data , developer made some model and put those values in that model and check the result that how much production should be enough so that complete product sellout in market so that company check . is there any trend or seasonal value from the data and check it and put same ratio in that model and obtain the result after that result they start the production.

So after getting the result then we are not sure that the result is correct then for this we have to create two diagram ACF and PACF and according the diagram if there is any seasonality then those diagram will lag with some values so now we will check how much multiple lag showing which is shows the seasonality.

5.3 Parameter Estimation

This issue another method in which we have to check all the parameter into the series so for a given series there will be many parameter so for this we have to select all the parameters which is correct for our analysis , suppose we have series which having many parameters like price , year , increase in price , average increment and price in particular time so we want to check the what will be the price of that product after some time so predict those value we have to choose some parameter which will be use full and will but this series in our model and find out the minimum parameter so for our requirement we need only two parameter which is year and price hike , those are the minimum parameter which will be used in our prediction and will try to find out the accurate price after some particular time.

5.3.1 Comparison Method

In this approach we will find out the method which we will used in our process. Now we have a lot of methods present in market so we can use any methods in from available but here our concern to select the method which is very appropriate for out prediction, from the all method which is present into the market, time series analysis is the best approach for solve the data which having with in some time of interval, so in this method we divide the data into the structure format and then try to check whether it stationary or not and for this operation we doesn't consider the whole series but take a small part of the series and perform our operation and getting that result on small amount to data now we check it for whole series and check weather its correct for the whole series after that verification we predict the future value.

So method selection is also play an important role in data science because method selection is also impact the result which we will get after perform the operation.

5.3.2 Standard Error Value for Parameter

After select the method for processing now we will check that parameter means minimum parameter which we have selected is correct or not because it's also important that selected parameter which we have selected in above method is correct or not, if it's not correct then whole result will produce the incorrect result so for this we have to take an stand error so that if will check firstly after selection the parameter.

5.3.3 Check Error for Value

Once the raw data was successfully processed and time-series process obtained then was made stationary. Then came the step to identify the process type of the time-series process to be used to predict the price of Copper price at Indian Stock Market. There are three different types of projects Auto-Regressive process, Moving-Average process and Auto-Regressive Integrated Moving-Average process. To identify the type of process to which the time-series process in-hand belongs required to plot two graphs. The two graphs plotted were Auto-correlation Function graph and Partial Auto-correlation Function graph. For identifying the time-series process as AR, MA or ARIMA process graphs were plotted and corresponding residual factor was noted down.

5.4 Error in Indices

In time series analyses we didn't check whole string in one time but get a small part of that series and for corresponding that part we check whether its stationary or not means draw a diagram for that part and check whether it's going to infinite or not, if it's going to infinite then part which have selected is not stationary so our main task to make stationary after that we will check the value of p which is find out in above equations if its value is less than 5% then that particular series is stationary and after that we get the result and check that result with remaining series and check the output sequence I similar to remaining series but result will not be same but there will be some difference in those values, so those difference is known as Error in Indices.

Mean value error (MVE): So in above if we will got any error from actual sequence and our prediction sequence then for this we will calculate the Means value error. So for this we will get the means of all the error present in our final graph , if we have some negative and positive error then they cancel out with each other and it will not give the correct result for error due to cancel out negative error value to positive.

Mean absolute value error (MAVE): This issue also very useful to predict in change the error so for this in above example we find out all the MVE (Mean value error) and take the average for that error check how we are getting average error on our prediction value. It's also play an important role to check error visibility.

Sum of squared value error (SSVE), Mean squared value error (MSVE): This issue the another approach to find out the error deviation so in this we do the average or sum of all the squared error which we are getting from prediction value and original values.

Percentage value error (PVE). In this we compute the error in the term of percentage so this firstly we will find out the error in maximum error and after getting the prediction value from the actual value and check what error we are getting so for this we will check what percentage error we are getting, so if the error we are getting more than 5% then this amount of error is not good to for this we have to manipulate our input again and change series and check error factor , again if we are getting more 5% then again manipulate the value till when we get less than 5% after that we will draw the diagram.

Mean percentage value error (MPVE). This issue also another type of error and in this error we take mean all the percentage error and check the result value. These percentage value comes from the series of small chunk of data which we are using to check the error.

Mean absolute percentage value error (MAPVE). In above method we percentage value error (PVE) and mean percentage value error (MPVE) we got the large number of positive or negative value so it’s not provide the fit chunk of series so that on which we can perform the operation, or due to positive and negative value, so they cancel out each other than to remove this we can use it mean absolute percentage value error (MAPVE)

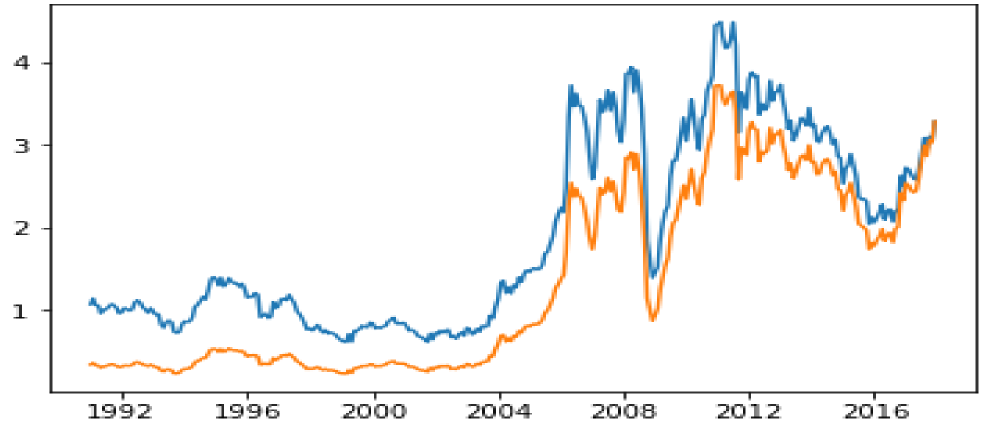
Automatic search to provide best result. To in this if we want to best result in case of error then if we are suing ARIMA auto regression integrated moving average it reduce the maximum probability of MAPVE or other type of error.

Chapter 6: Conclusion

6.1 Conclusive Results

Firstly we have a data series and our main task is that to make that series as time series so we have checked all the data and check weather its follow all the condition to make a stable series if it full fill all the condition then our main task to make that series more stable or stationary so we have use different techniques to make it as stationary.

Firstly we have a data series and our main task is that to make that series as time series so we have checked all the data and check weather its follow all the condition to make a stable series if it full fill all the condition then our main task to make that series more stable or stationary so we have use different techniques to make it as stationary.



Now we have all the value for forecast the value then we will draw or plat a diagram with forecast values and check difference and tread of the prediction value with the original value if difference is small then follow same trend then our predication value is correct then we will predict more data.

Above is the resultant graph which we get after perform many operation on the given data set so in that graph we have two lines one is blue and other is yellow, blue line is the actual result and yellow line which we got as future prediction so from above graph we can say that both the line having very less difference so we can say that over predications correct with some offset value.

6.2 Conclusive Summary

Time series play an important role in today's data analysis. In our data we divide data series but type of variable is using, if its denoted by single variable then no need to check those series from time series analysis because for this type of series we can easily draw the diagram and get the result what will be the next output, but if our data series having multi variable then time series is important for such type data so there are many system which are using time series analysis like flight reservation system also using this method for prediction and other important useful area is stock market , this market also using this method for prediction.

Additional advanced time series techniques include Fourier transformation and decomposition for any value or series autocorrelation analysis which will help to check any seasonality or trend related things. The main reason to use time series analysis to find out the pattern or behavior of the data series with corresponding change or movement of the variable which can be single variable or multiple variable over time.. Also there is any important factor which can affect our patterns is trends and seasonality, both factor play major role for predication the value our structure data set.

6.3 Future Work

In this method data which we are found is not so accurate so in future we will try to achieve accuracy or try to achieve accuracy with existing data. And try to put some better approx. to check trend and seasonality related data.

REFERENCES

[1] Prediction of Rupiah Against US Dollar by Using ARIMA, Adiba Qonita, Annas Gading Pertiwi, Triyanna Widiyaningtyas,Electrical Engineering Department,Universitas Negeri Malang,Malang, Indonesia

[2] Short-term Traffic Flow Prediction Using a Methodology Based on ARIMA and RBF-ANN,Kui-lin Li,Chun-jie Zhai,Jian-min Xu,School of Automatic Science and Engineering South China University of Technology Guangzhou,China.

[3] ARIMA Implementation to Predict the Amount of Antiseptic Medicine Usage in Veterinary Hospital, Hans Pratyaksa, Adhistya Erna Permanasari, Silmi Fauziati, Ida Fitriana,Department of Electrical Engineering and Information Technology, Department of Pharmacology Universitas Gadjah Mada ,Indonesia.

[4] Time series forecasting using improved ARIMA, Soheila Mehrmolaei Computer Engineering Qazvin Branch, Islamic Azad University Qazvin, Iran.

[5] Airline passenger forecasting using neural networks and Box–Jenkins, S.M.T. Fatemi Ghomi and K. Forghani Department of Industrial Engineering Amirkabir University of Technology Tehran, Iran.

[6] Application and analysis of forecasting stock price index based on combination of ARIMA model and BP neural network, Yulin Du, School of Management ,Fudan University ,Shanghai,China.

[7] Forecasting Method of Aero-Material Consumption Rate Based on Seasonal ARIMA Model, Yanming Yang, Chenyu Liu, Feng Guo, Qingdao Campus, Naval Aeronautical University, Qingdao 266041, China.

[8] Forecasting of Raw Material Needed for Plastic Products Based in Income Data Using ARIMA Method, Baihaqi Siregar, Erna Budhiarti Nababan, Alexander, Yap, Ulfi Andayani, Department of Information Technology, University of Sumatera Utara, Medan, Indonesia.

[9] Effects of Changes in Earned Income Tax Credit: Time-series Analyses of Washington DC.

[10] Modelling and Forecasting of Sri Lankan Fishery Exports through Autoregressive Integrated Moving Average Methodology.

[11] Evaluation of Forecasts Performance of ARIMA-GARCH-type Models in the Light of Outliers.