A PROJECT ON

5 YEARS SALES FORCAST BASED ON AUTO REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL

BY

PAVAN M SUNDER

ABSTRACT

The aim of this project is to obtain 5 years future sales forecast for a time series data based on **A**uto **R**egressive **I**ntegrated **M**oving **A**verage **(ARIMA)** model.

The data set used for this project is monthly number of sales of shampoo over a 3-year period taken from Kaggle website.

A stationary data is generally desirable to work on as it facilitates easier and reliable implementation of predictive models. Stationarity means that the statistical properties of a process generating a time series data do not change over time.

However, most often the real-life data will be non-stationary and we might have to apply a certain transformation to make the data stationary. The data set used for this project was initially non – stationary with increasing trend and seasonal variations. A transformation was applied to make the data stationary.

ARIMA model is then fitted on transformed data set. Then future 5 years future sales forecast was obtain based on fitted model.

CONTENTS

1. INTRODUCTION

There are so many prediction problems that involve a time component. In this project I have considered a time series dataset for analysis and prediction. A time series data is a series of data points indexed (or listed or graphed) in time order. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series forecasting** is the use of a model to predict future values based on previously observed values.

Autoregressive Integrated Moving Average (ARIMA) model is one of the models we can deploy to better understand the time series data or to predict future points in the series (forecasting). ARIMA is a generalization of an Auto Regressive (AR) Moving Average (MA) model. ARIMA models can be applied in some cases where data show evidence of non-stationarity. An initial differencing step (corresponding to the "Integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

Non-seasonal ARIMA models are generally denoted ARIMA(p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model.

In this project I have used ARIMA model to forecast 5 years sales numbers for a non-stationary shampoo sales time series dataset.

2. DATA

The data set used for this project is monthly number of sales of shampoo over a 3-year period taken from Kaggle website. The units are a sales count and there are 36 observations. The original dataset is credited to Makridakis, Wheelwright, and Hyndman (1998)

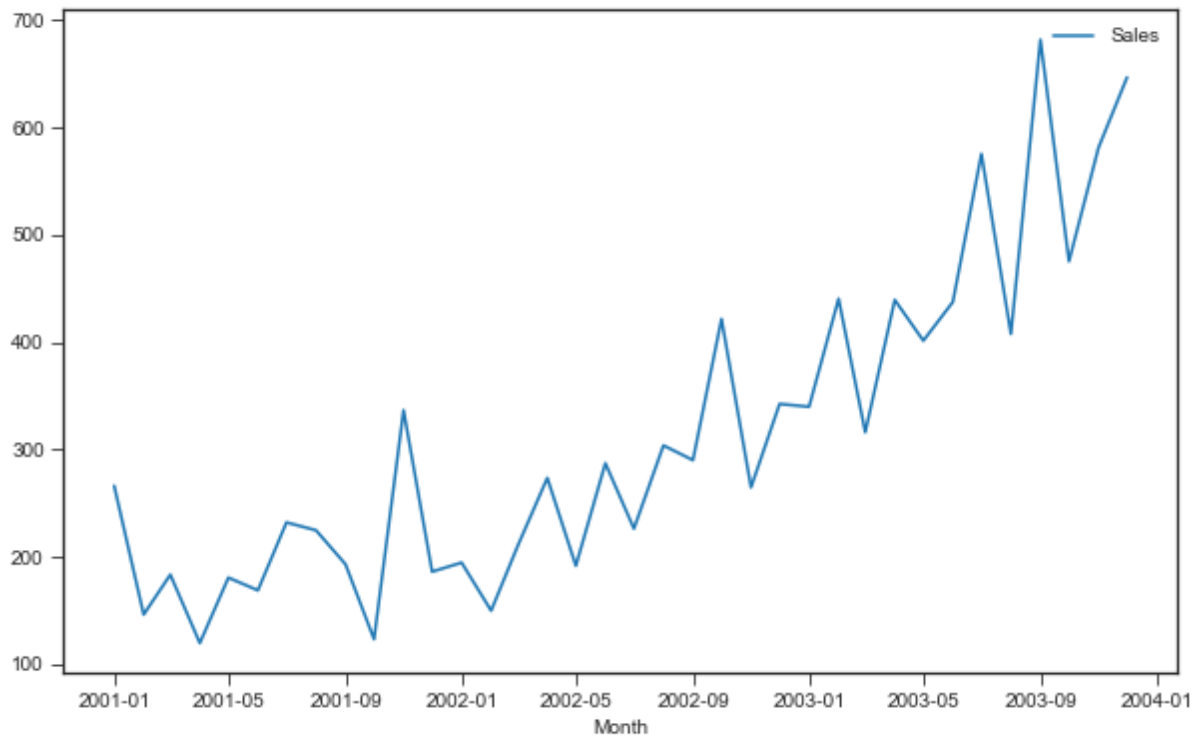
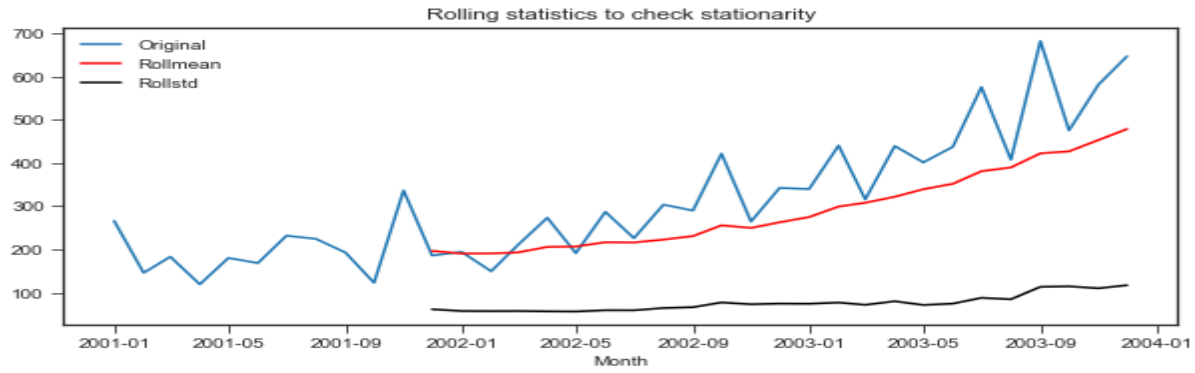


Figure 1. Original shampoo sales data

The data apparently shows non – stationarity with increasing trend and seasonal variations.

The non – stationarity can further be confirmed by using rolling statistics plot and augmented dickey fuller test as shown in Figure 2.




```
Results of Dickey-Fuller Test:
Test Statistic                  -0.945353
p-value                          0.772671
#Lags Used                      10.000000
Number of Observations Used     25.000000
Critical Value (1%)             -3.723863
Critical Value (5%)             -2.986489
Critical Value (10%)            -2.632800
dtype: float64
Testing test statistic (-0.9453530717367251) < 1% confidence interval
(-3.7238633119999998)?: False
```

Figure 2. Rolling statistics plot and dickey fuller test results of original dataset.

The illustration in Figure 2 clearly shows increasing trend (positive slope of rolling mean denoted in 'red'). Also, the test statistic according to Dickey Fuller test is greater than the critical values. There is also higher p value. Thus, we can say that the data is non-stationary.

3. DATA PREPROCESSING

The non – stationary data needs to be transformed to make it stationary before applying ARIMA model.

The following steps illustrates transformation of data in order to render it stationary.

1. Taking logarithm of the original time series data. Let '*ts*' denote original time series data and '*ts_trans_log*' denote log transform of original time series data ('*ts*')

$$ts\_trans\_log = log(ts) \quad (1)$$

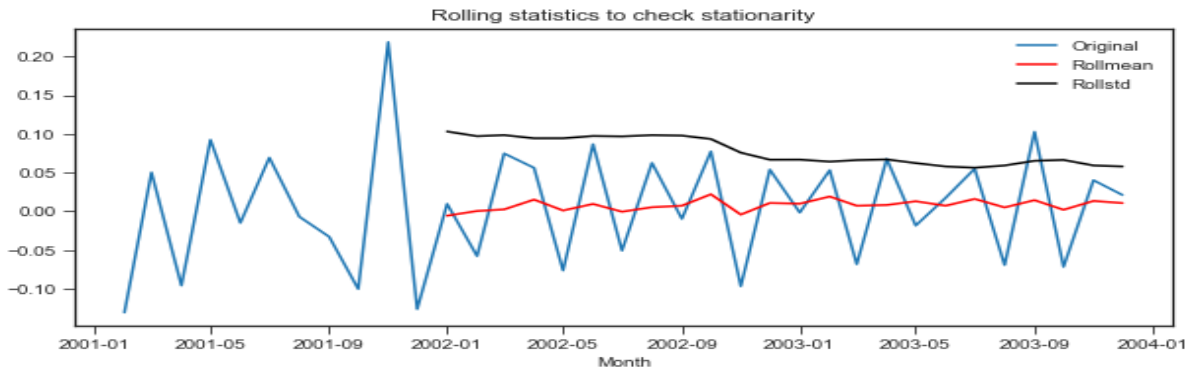
2. Taking square root of log transformed data (‘*ts_trans_log’)* denoted as ‘*ts_trans*’

$$ts\_trans = \sqrt[2]{ts\_trans\_log} \quad (2)$$

Taking 1st order difference of square root transformed data (‘*ts_trans*’) denoted as ‘*ts_trans_diff*’

$$ts\_trans\_diff = diff(ts\_trans) \quad (3)$$

The stationarity of the transformed data can further be validated and confirmed using the rolling statistics plot and Dickey Fuller test as described in Figure 3.




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.960377
p-value                          0.001633
#Lags Used                       9.000000
Number of Observations Used     25.000000
Critical Value (1%)             -3.723863
Critical Value (5%)             -2.986489
Critical Value (10%)            -2.632800
dtype: float64
Testing test statistic (-3.960376979624248) < 1% confidence interval
(-3.7238633119999998)?: True
```

Figure 3. Rolling statistics plot and dickey fuller test results of transformed dataset.

The illustration in Figure 3 shows constant mean (denoted in ‘red’). Also, the test statistic according to Dickey Fuller test is less than 1% critical value. There is also low p value. Thus, we can say with 99% confidence that the data is stationary.

Now we can fit the ARIMA model to the transformed data as it is stationary.

4. ARIMA MODEL

In order to fit the ARIMA model to the transformed data we need to determine optimal parameters i.e. (p, q, d) values.

We have seen that the data is rendered stationary by taking 1st order difference. Therefore, the optimal d value shall be 1.

Further we need to determine p and q values. One of the ways to find p and q values is by plotting partial autocorrelation and auto correlation plots respectively as illustrated below.

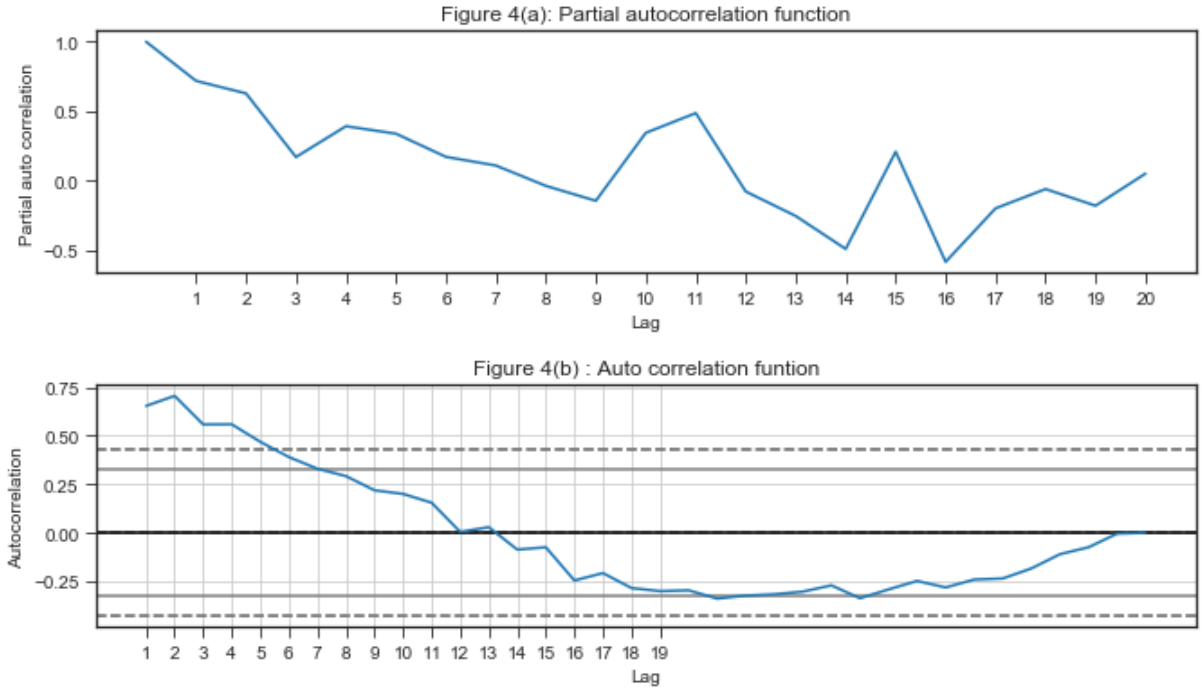


Figure 4. (a) Partial auto correlation plot; (b) Auto correlation plot

Figure 4(a) represents Partial auto correlation plot from which we can determine optimum p value. From the figure we observe that the plot crosses the upper confidence interval for the first time at approximately 1, hence, p shall be 1.

Figure 4(b) represents Auto correlation plot from which we can determine optimum q value. From the figure we observe that the plot crosses the upper confidence interval for the first time between 5 and 6, hence, q shall be 5 or 6.

Based on above insights, the ARIMA model is fitted to the time series data by taking p =1, d = 1, q = 6. Figure 5 illustrates ARIMA model fitted to the transformed data.

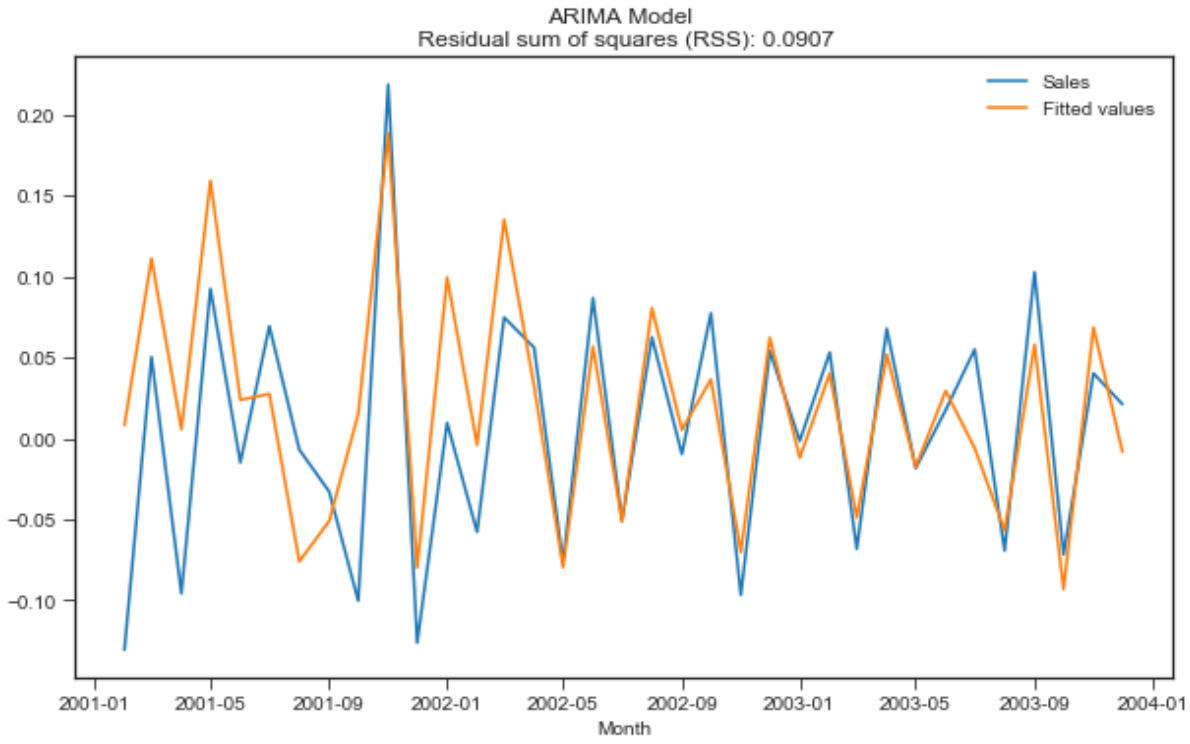


Figure 5. ARIMA model fitted to the transformed data.

From Figure 5 we can observe that the model has fairly fitted the data well indicated by the low Residual sum of squares (RSS) value of 0.0907. The model has also captured the seasonal variations present in the input data.

The input data for the ARIMA model is a transformed data from which the trend present in the original time series was removed to make it stationary. However, it is necessary to bring back the fitted values to original form inorder to incorporate the trend that was removed and make reliable predictions.

This can be achieved by the following steps:

1. First get the fitted values and store it as series. We will notice that the first month will be missing because lag of 1(shift) was taken to carry out 1st order difference.
2. Convert differencing to log scale: find the cumulative sum and add it to a new series with a base value (here the first-month value of the log series).
3. Take the exponent of the series from above (anti-log) and square it which will be the predicted values of the time series forecast model as illustrated in Figure 6

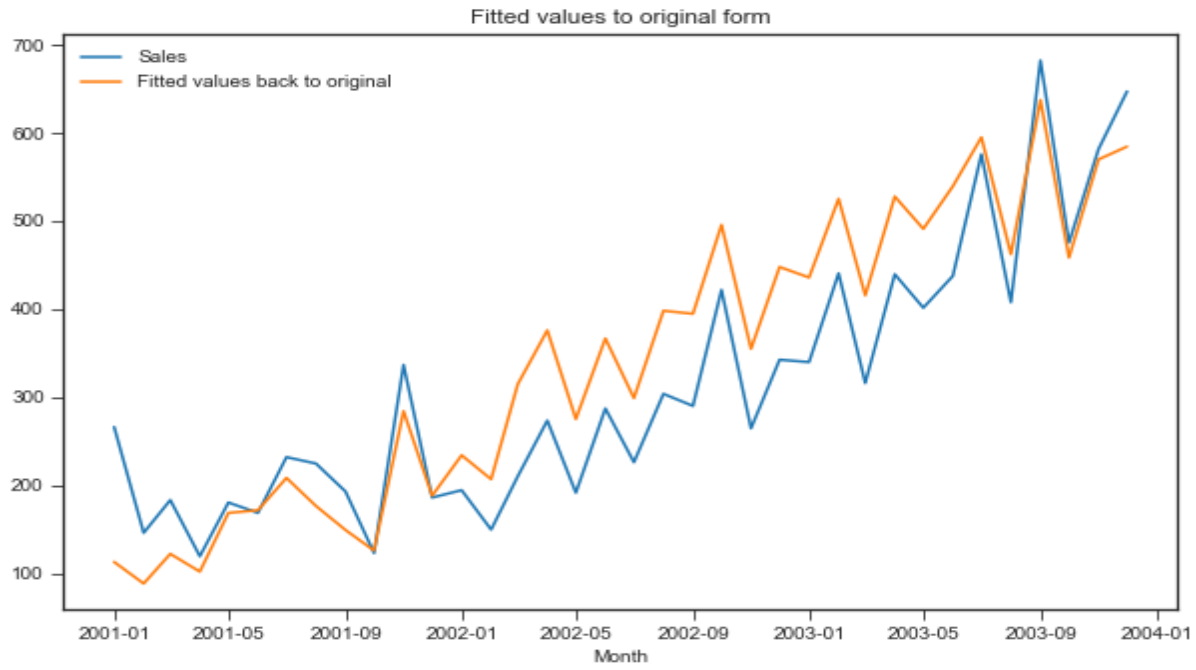


Figure 6. ARIMA model fitted values to original form.

4. FORECAST FOR 5 YEARS

5 years future sales forecast can be obtained from the fitted values of the ARIMA model as illustrated in Figure 7.

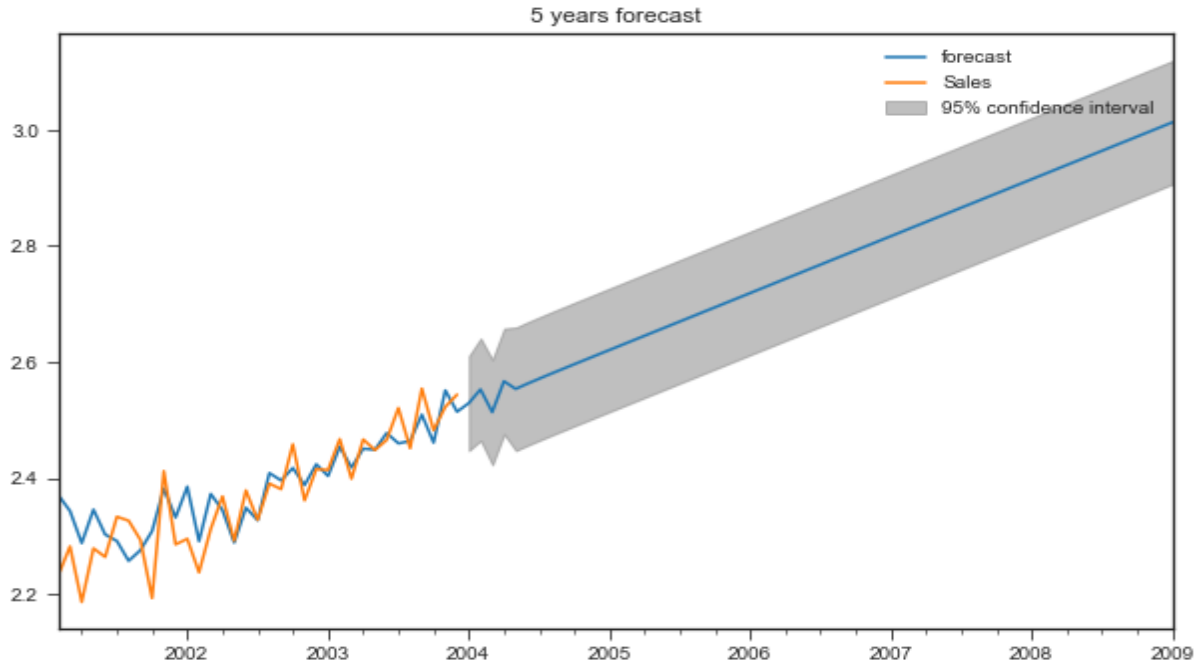


Figure 7. 5 years sales forecast.

5. CONCLUSION

I have tried to implement ARIMA predictive model for a non-stationary time series. The parameters for the model was taken as p=1, d=1 and q=6 based on rolling statistics plots, Dickey Fuller test, auto correlation and partial auto correlation plots.

The model was able to fit the data fairly well with the selected parameters capturing the trend and seasonal variations. I was also able to get the 5 years forecast based on ARIMA model.

The complete python code for this project is available in my github repository. Follow this link to access the code.