Project Report

SAMSUNG

R&D Institute, Bangalore

MAGPIE

#NotJustAnyIntern

Long Call Timer

Team- Network Access and Core R&D

Submitted by-

Abhishek Raj Permani

MAGPIE Intern’20

IIT(ISM) Dhanbad

Index

Problem Statement

Most of service providers have a threshold value of each call, once the call duration exceeds this threshold value, the call automatically gets disconnected, and this value cannot be very small (because this will end call very frequently, which will leave users unsatisfied) also cannot be very large(because this will increase the load on network lines), it should be optimal.

We have to find that value of time after which the user has to redial again in order to continue. Also, a fix percentage of users must be satisfied with that time value.

Satisfied Percentage of users- is the percentage of total users, who ends the call before that end call time value, which we have to calculate.

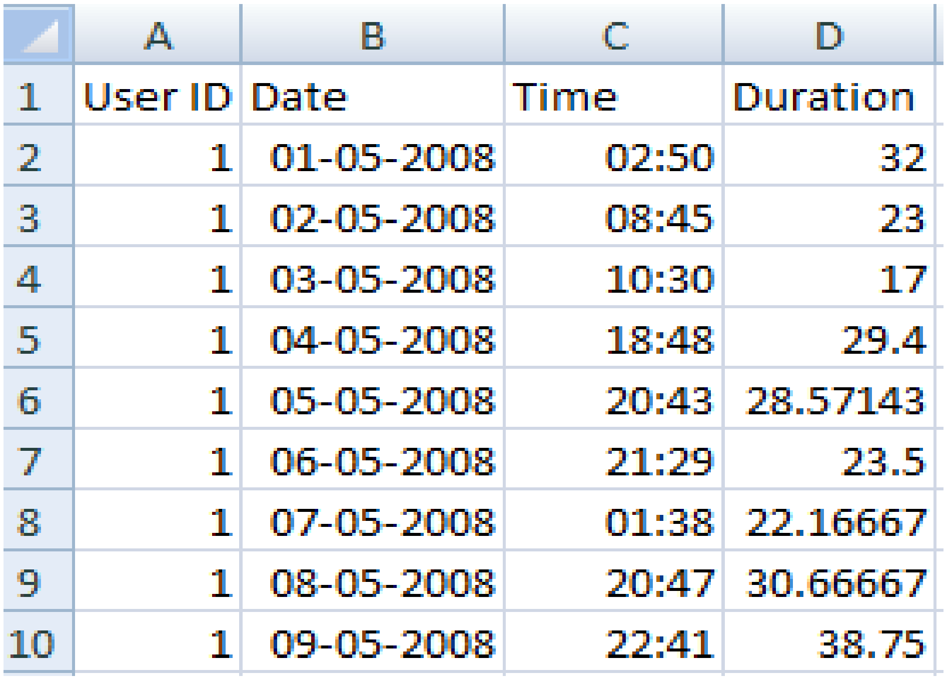
Assumption

- We have assumed that many users can simultaneously make phone calls over the network.

Approach

1) Preparing Dataset-

- Dataset of 30 users for 25 Days has been taken in consideration
- Average call duration of each user on each single day has been taken into account.
- Dataset is generated artificially using Python.



	A	B	C	D
1	User ID	Date	Time	Duration
2	1	01-05-2008	02:50	32
3	1	02-05-2008	08:45	23
4	1	03-05-2008	10:30	17
5	1	04-05-2008	18:48	29.4
6	1	05-05-2008	20:43	28.57143
7	1	06-05-2008	21:29	23.5
8	1	07-05-2008	01:38	22.16667
9	1	08-05-2008	20:47	30.66667
10	1	09-05-2008	22:41	38.75

Part of Dataset for User with User ID-1

2) Introduction to ARIMA Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data.
It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.

This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
  Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

- **p**: The number of lag observations included in the model, also called the lag order.
- **d**: The number of times that the raw observations are differenced, also called the degree of differencing.

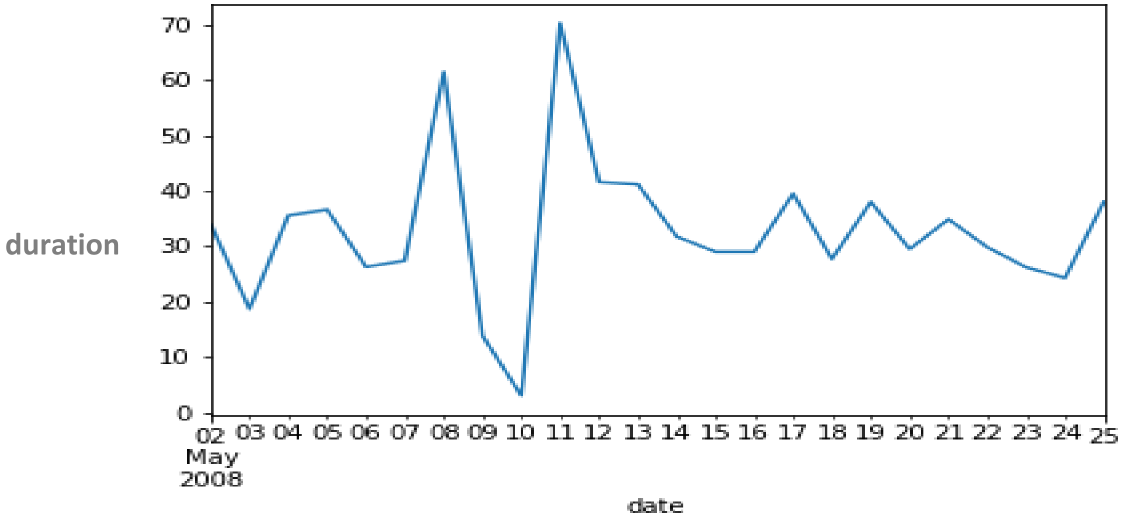
- **q**: The size of the moving average window, also called the order of moving average.
  A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model.

A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious, but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

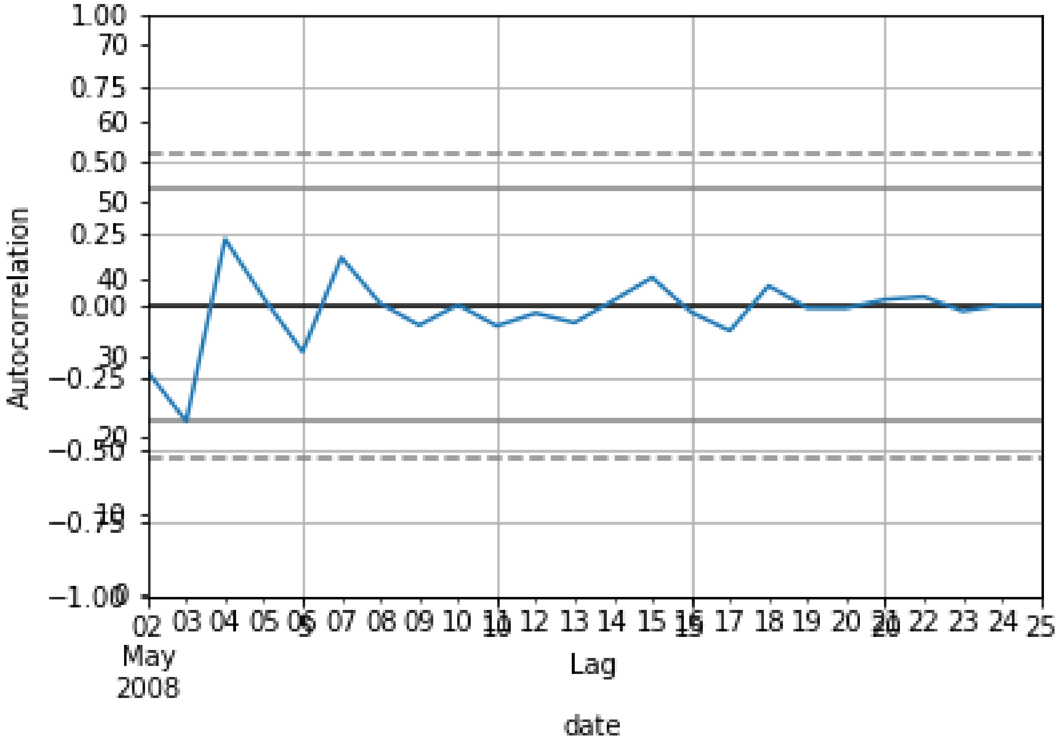
3) Working on prepared Dataset-

This is the plot of a particular user's call duration versus date of that call.



Plot of Data of User ID-1 (Duration vs Date)

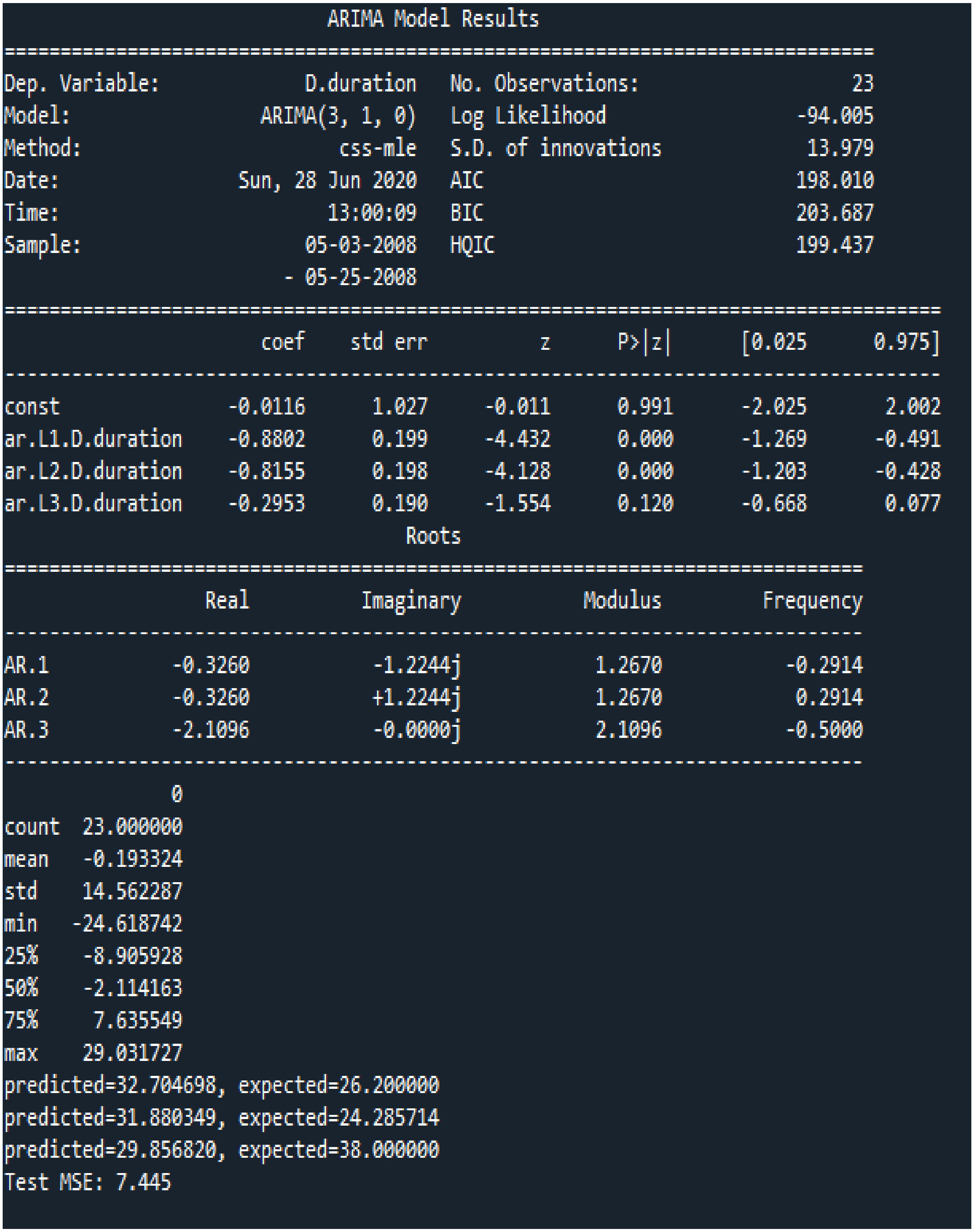
Let’s also take a quick look at an autocorrelation plot of the time series.



Autocorrelation Plot of User ID-1 Dataset

When fitting the model, a lot of debug information is provided about the fit of the linear regression model. We can turn this off by setting the *disp* argument to 0.

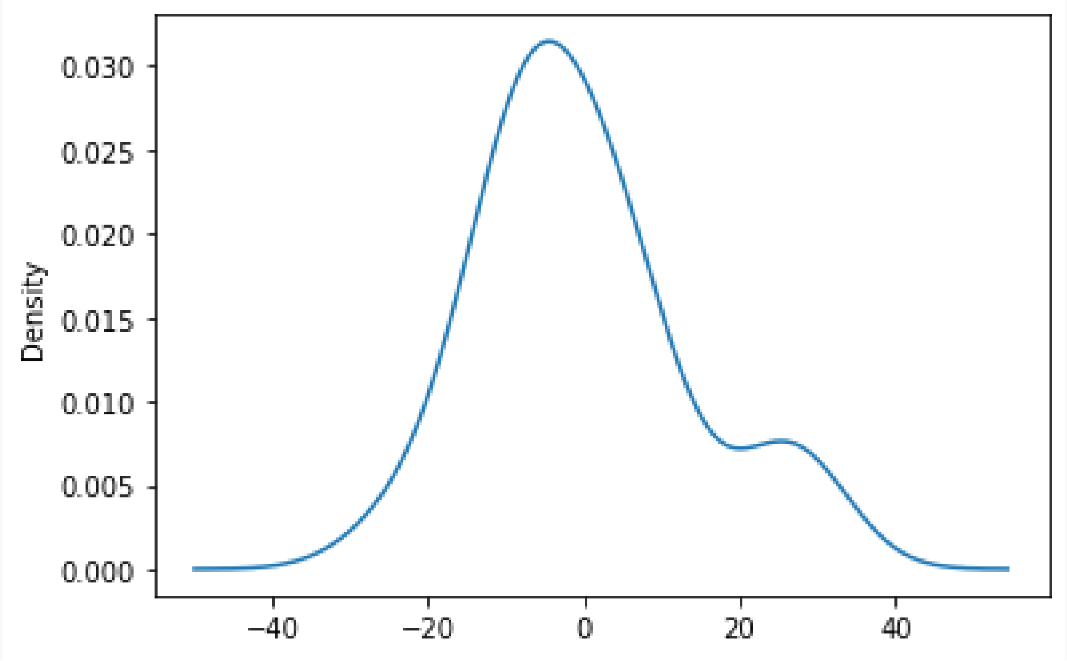
This summarizes the coefficient values used as well as the skill of the fit on the on the in-sample observations.



```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.duration   No. Observations:                   23
Model:                 ARIMA(3, 1, 0)   Log Likelihood                 -94.005
Method:                       css-mle   S.D. of innovations             13.979
Date:                Sun, 28 Jun 2020   AIC                            198.010
Time:                        13:00:09   BIC                            203.687
Sample:                    05-03-2008   HQIC                           199.437
                         - 05-25-2008
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.0116      1.027     -0.011      0.991      -2.025       2.002
ar.L1.D.duration     -0.8802      0.199     -4.432      0.000      -1.269      -0.491
ar.L2.D.duration     -0.8155      0.198     -4.128      0.000      -1.203      -0.428
ar.L3.D.duration     -0.2953      0.190     -1.554      0.120      -0.668       0.077
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.3260           -1.2244j            1.2670           -0.2914
AR.2           -0.3260           +1.2244j            1.2670            0.2914
AR.3           -2.1096           -0.0000j            2.1096           -0.5000
-----------------------------------------------------------------------------
               0
count  23.000000
mean   -0.193324
std    14.562287
min   -24.618742
25%    -8.905928
50%    -2.114163
75%     7.635549
max    29.031727
predicted=32.704698, expected=26.200000
predicted=31.880349, expected=24.285714
predicted=29.856820, expected=38.000000
Test MSE: 7.445
```

summary of the fit model

Density plot of the residual error values, suggesting the errors are Gaussian, but may not be centered on zero. The distribution of the residual errors is displayed. The results show that indeed there is a bias in the prediction (a non-zero mean in the residuals).



ARIMA Fit Residual Error Density Plot

Mathematics of ARIMA Model-

A non-seasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

The forecasting equation is constructed as follows. First, let y denote the d^th^ difference of Y, which means:

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

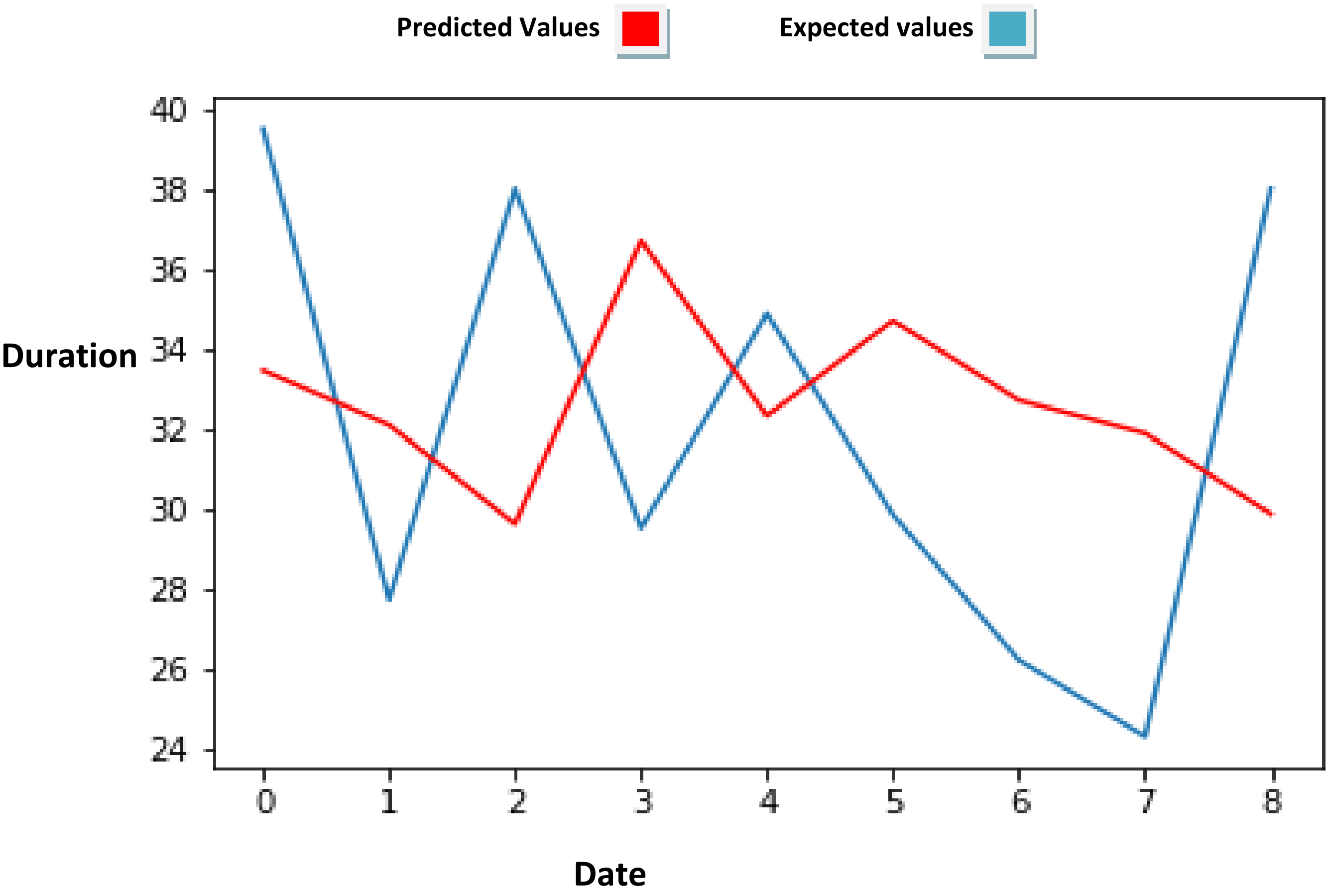
Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the *first-difference-of-the-first difference*, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.

In terms of y, the general forecasting equation is:

$\hat{y}_t = \mu + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$

After performing above calculations, and observing the above dataset, we will now forecast the values of call duration of a particular user shown below.

A rolling forecast is required given the dependence on observations in prior time steps for differencing and the AR model. A crude way to perform this rolling forecast is to re-create the ARIMA model after each new observation is received.

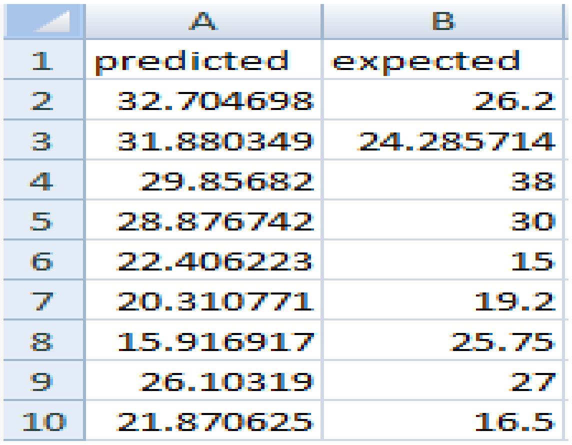


ARIMA Rolling Forecast Line Plot

(A line plot is created showing the expected values (blue) compared to the rolling forecast predictions (red). We can see the values show some trend and are in the correct scale.)

Manually keep track of all observations in a list called history that is seeded with the training data and to which new observations are appended each iteration.

Below is the list of predicted and expected values of call duration average of all 30 users taken into account, these values will be used for calculating - Root mean squared error (RMSE) value for dataset.



	A	B
1	predicted	expected
2	32.704698	26.2
3	31.880349	24.285714
4	29.85682	38
5	28.876742	30
6	22.406223	15
7	20.310771	19.2
8	15.916917	25.75
9	26.10319	27
10	21.870625	16.5

Expected values compared to the rolling forecast predictions

The forecast errors are on the same scale as the data. Accuracy measures that are based only on e_t are therefore scale-dependent and cannot be used to make comparisons between series that involve different units.
The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

Mean absolute error: MAE=mean(|e_t|)

Root mean squared error: RMSE=√$\boxed{\text{mean}(e_t^2)}$

(where e_t is the error term)

- Mean absolute error: MAE=mean(|et|),Root mean squared error: RMSE=mean(et2).When comparing forecast methods applied to a single time series, or to several time series with the same units, the MAE is popular as it is easy to both understand and compute. A forecast method that minimises the MAE will lead to forecasts of the median, while minimising the RMSE will lead to forecasts of the mean. Consequently, the RMSE is also widely used, despite being more difficult to interpret.

Limitations

Time Series data has few properties like seasonality, trend, which leads to a better fit and lesser value of RMSE value, thus a better forecast. The above Dataset is generated artificially using python, so this randomness sometimes does not guarantees the properties of time series in data set.

Conclusion

The RMSE value for the above Dataset comes out to be= 15.879966352

Once we have obtained the forecasted data using above ARIMA Model, we can calculate the threshold value. For the above dataset taken into consideration, the threshold time value to satisfy 95% of the users comes out to be = 55.09 minutes.