Priya Diwakar
200205361

IOT Analytics

Project 3 Forecasting

1. Task 1 Check for Stationarity

On visually inspecting the time series data plot in Figure 1, it is observed that it is stationary. There are no predictable patterns. There doesn't seem to be any of the features of non-stationarity such as trends, variable variance or seasonality. Based on these observations no transformations are applied to the time series.

Figure 2 shows the training and testing set split, where training set has 1500 observations and testing set has 500 observations.

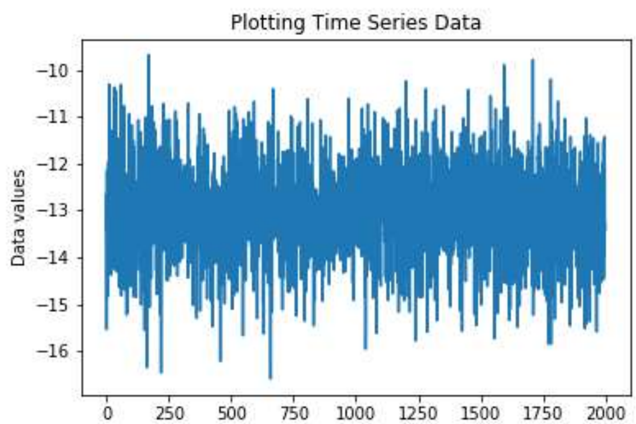

Fig. 1 Plot to Check Stationarity

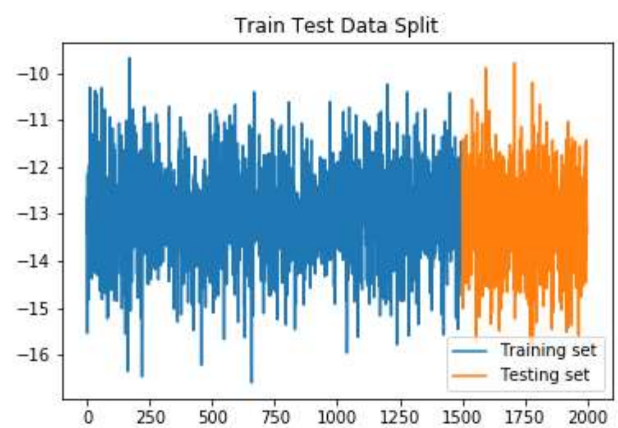

Fig. 2 Train Test Split

2. Task 2 Fit a Simple Moving Average Model on the training set

A Simple Moving Average Model as shown in Eqn. (1), is fit on the training data for different values of k ranging from 2 to 290. The root mean squared error (RMSE) is computed between original and predicted values. The final model is fit for the k value which yields the lowest RMSE value.

$$s_t = \left(\frac{1}{k}\right)\sum_{i=t-k}^{t-1} x_i \qquad \text{............. Eqn. (1)}$$

Figure 3 shows the RMSE vs k plot. From the plot it is observed that at k equal to 288 the RMSE is lowest with a value of 0.9713. The RMSE value initially decreases quite rapidly and then slowly converges to the lowest value.

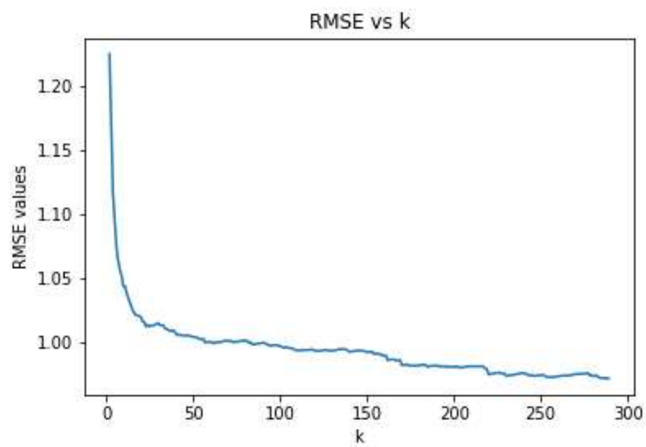

Fig. 3 RMSE vs k

Figure 4 shows the original (training set) and predicted values for the time series data. Visually the predicated values are not very accurate but the RMSE value is quite small.

Fig. 4 Original against Predicted values

3. <u>Task 3 Fit an Exponential Smoothing Model on the training set</u>

An Exponential Smoothing Model as shown in Eqn. (2), is fit on the training data for different values of alpha ranging from 0.1 to 0.9 with a step size of 0.1. The root mean squared error (RMSE) is computed between original and predicted values. The final model is fit for the alpha value which yields the lowest RMSE value.

$$s_t = ax_{t-1} + (1\text{-}a)s_{t-1} \quad \text{............. Eqn. (2)}$$

Figure 5 shows the RMSE vs alpha plot. From the plot it is observed that at alpha equal to 0.3 the RMSE is lowest with a value of 1.2014. The RMSE value initially decreases till alpha equal to 0.3 and then increases till alpha equal to 0.9.

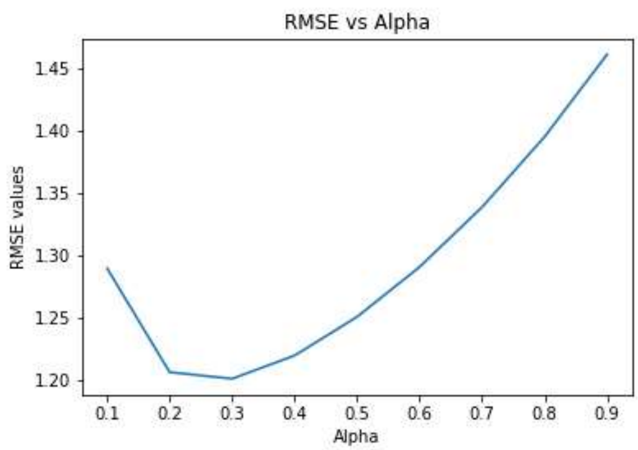

Fig. 5 RMSE vs alpha

Figure 6 shows the original (training set) and predicted values for the time series data. Visually the predicated values are not perfectly accurate but the RMSE value is reasonably small though greater than the Simple Moving Average model.

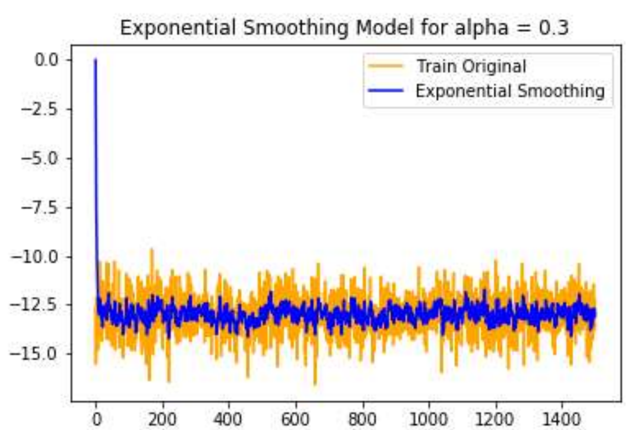

Fig. 6 Original against Predicted values

4. <u>Task 4 Fit an Autoregressive AR(p) Model on the training set</u>

An Autoregressive AR(p) Model as shown in Eqn. (3), is fit on the training data. The order p for the AR model is selected by observing the Partial Autocorrelation (PACF) plot of the data. Figure 7 shows the PACF plot. From the plot the order for the AR model is selected as 1, as the first zero coefficient is obtained at lag equal to 2.

$$X_t = \delta + a_1X_{t-1} + a_2X_{t-2} + \ldots + a_pX_{t-p} + \varepsilon_t \qquad \text{............. Eqn. (3)}$$

The estimated parameters for the AR(1) model are

$$\delta = -13.012198 \text{ and } a_1 = -0.137609$$

The root mean squared error (RMSE) is computed between original and predicted values and is found to be 0.988.

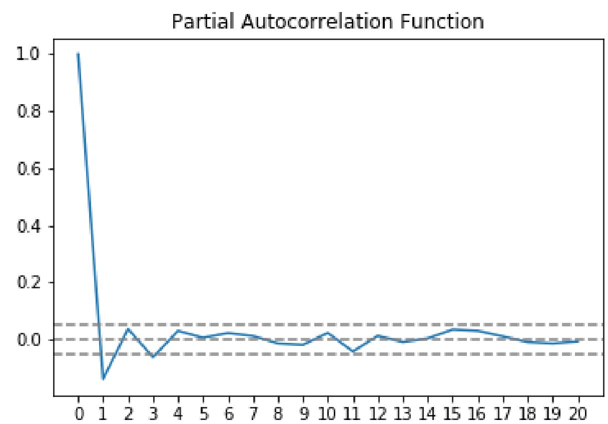

Fig. 7 PACF Plot

Figure 8 shows the original (training set) and predicted values for the time series data. Visually the predicated values are not very accurate but the RMSE value is quite small though slightly greater than the Simple Moving Average model but less than the Exponential Smoothing model.

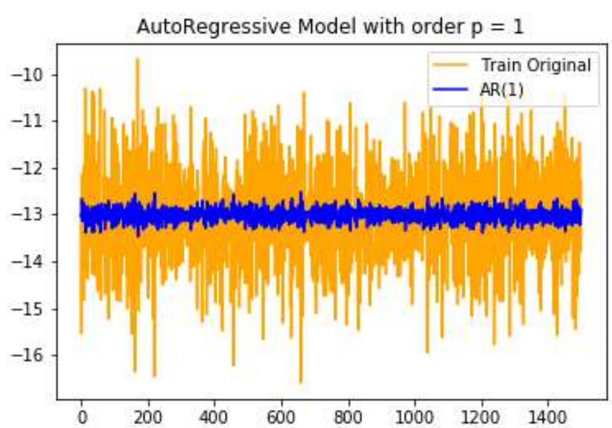

Fig. 8 Original against Predicted values

Residual Analysis:

a. Q-Q plot and Chi-squared test

1. Q-Q plot:

   The pdf of residuals against the pdf of $N(0,s^2)$ is shown in the Q-Q plot in Figure 9. The two distributions are almost equal, though the left tail of residuals strays off a bit from the line as the quantile is a slightly smaller value.

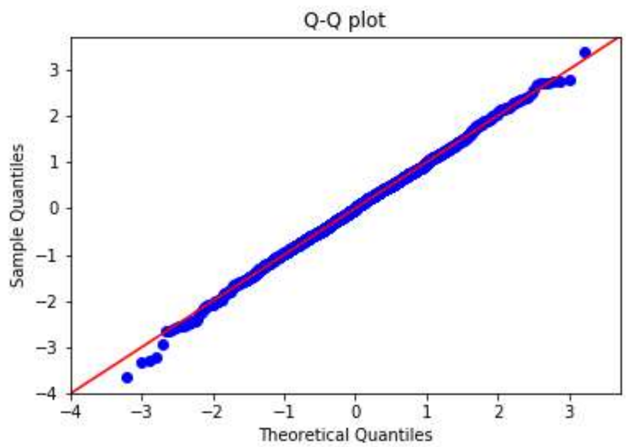

Fig. 9 Q-Q Plot

2. Chi-Squared Test:

   The chi-squared test gives a p value of 0.3607 so we accept the null hypothesis and say the pdf of the residual is a normal distribution at all confidence levels 90%, 95% and 99%. Figure 10 shows the histogram plot of the residuals. Visually the histogram plot does resemble the gaussian pdf.

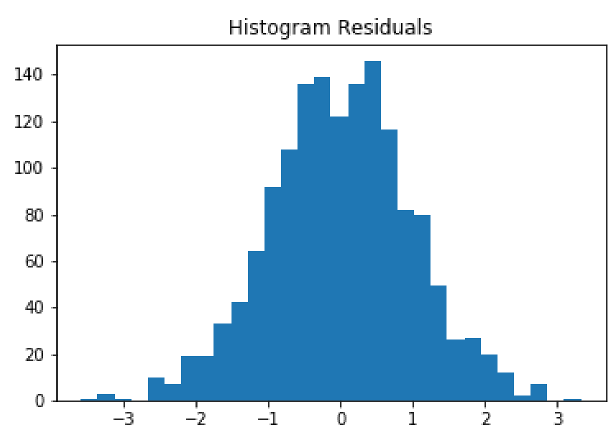

Fig. 10 Histogram Plot

b. Residual Scatter Plot:
In the scatter plot shown in Figure 11 the residuals are random and no correlation trends are observed.

Based on the residual analysis the model is a good fit as the assumptions for the residuals hold true.

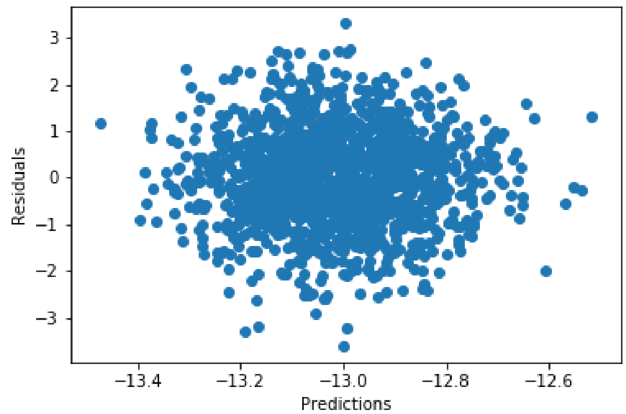

Fig. 11 Residual Scatter Plot

5. Task 5 Comparison of the three models using the testing set

1. Simple Moving Average Model
The root mean squared error (RMSE) is computed between test and predicted values and is found to be 1.0533.
Figure 12 shows the original (test set) and predicted values for the time series data. Visually the predicated values are not very accurate but the RMSE value is reasonably small.

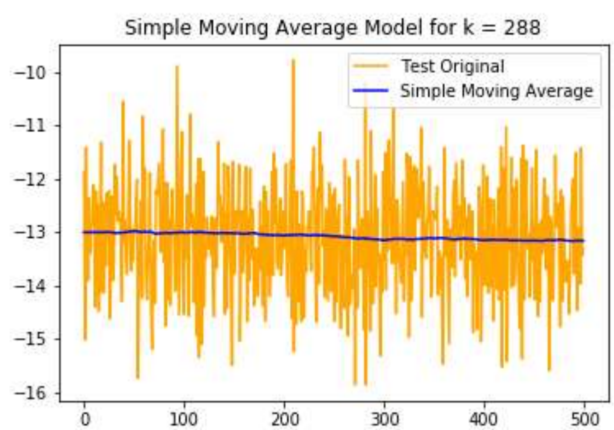

Fig. 12 Original against Predicted values

2. Exponential Smoothing Model
   The root mean squared error (RMSE) is computed between test and predicted values and is found to be 1.1746.
   Figure 13 shows the original (test set) and predicted values for the time series data. Visually the predicated values are not completely accurate but the RMSE value is reasonably small though slightly greater than the Simple Moving Average model.

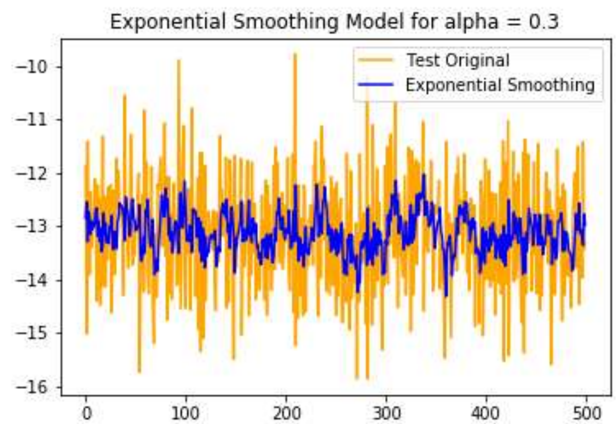

Fig. 13 Original against Predicted values

3. Autoregressive AR(1) Model
   The root mean squared error (RMSE) is computed between test and predicted values and is found to be 1.0249.
   Figure 14 shows the original (test set) and predicted values for the time series data. Visually the predicated values are not completely accurate but the RMSE value is the smallest among the three models.

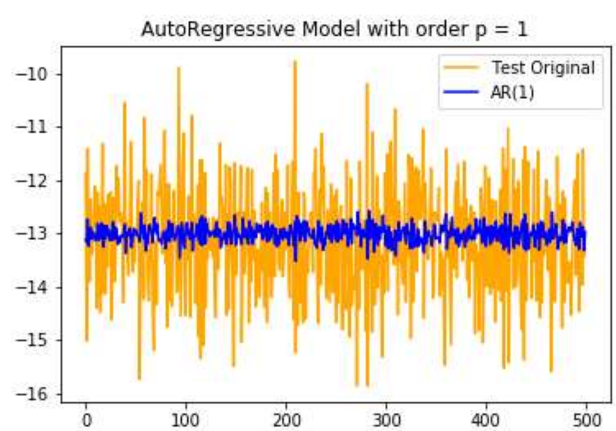

Fig. 14 Original against Predicted values

Based on the RMSE values obtained for each model for the testing set, the Autoregressive AR(1) Model would be the best choice of model for this time series.