**Practical Machine Learning Exercise – Solution**

First step in any machine learning problem is to read data and ensure all the data types are proper and look for missing data.

After importing the sales and traffic data, I checked to see if the data is present for all dates. Both Sales and Traffic dataset had some missing dates. So, I wrote a function to find out the dates that were missing. Sales and Traffic data had missing rows during Christmas, Easter days and Thanksgiving days. So those dates were added to the dataset with value of 0.

**Explanatory Data Analysis:**

EDA was performed to understand the relationship between Sales and Traffic. From the plots, It was evident that whenever there were peaks in Sales there were subsequent peaks in Traffic Data as well. Further analysis showed that these peaks occurred between Thanksgiving and Christmas. Two new columns called Leave and Holiday Season were created. The holiday season has a value ‘1’ for 20 days before Christmas. The Easter and Christmas days were marked as ‘1’ in ‘Leave’ column. Analysis was done to explore if sales happened when there was no traffic and vice versa.

**Feature Engineering and Data Augmentation:**

To further understand the data, sales and traffic values were discretized and bins were created. Bins were created in such way that all bins had enough observations in them. The Hour, Minutes, Day, Day of Week, Month and Year column information were extracted from Timestamp field and separate column were created. Sales and Traffic were analyzed by hours. Further one more column called Quarter was created and Sales and Traffic were analyzed for each quarter.

**Model 1: Random Forest Classification**

The Sales Range and Traffic Range columns were treated as categorical columns. Hence, it becomes a multi classification problem. Random Forest model was chosen. The model predicted very closely to actual values. Before building the model, I performed granger causality test to determine the relationship between Sales and Traffic. It helps to verify If Sales causes Traffic, that is if change in Sales is followed by change in Traffic, then Traffic can be help in predicting Sales. So, to investigate whether if Traffic causes Sales, I set null hypothesis as: “Traffic does not cause Sales” i.e. the null hypothesis is there is no causation between Sales and Traffic. If the null hypothesis is rejected, then there is evidence that granger causality exists. i.e. Traffic causes Sales. I set the critical value as 0.05. The lags in which the p-value obtained from granger causality between Sales and Traffic were below 0.05. So, the null hypothesis was accepted and the model was built by including Sales for Traffic prediction and vice versa.

**Model 2: Time Series – ARIMA**

The data must be in Stationary Time Series format for forecasting. So, I converted the data to Time Series format with frequency based on hour. Summed the sales and traffic based on hour and made it as index. Next the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were plotted to identity if the time series is in Stationary or Non-Stationary format. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are measure of association between current and past series values and indicate which past series values are most useful in the prediction of future values. If ACF dies out slowly and PACF drops to zero suddenly after lag 1 and if correlation is different from zero then data is non-stationary. If ACF has positive values for higher number of lags and it does not drop-off to zero quickly after few lags, then it means it has a high trend and is highly non-stationary and requires higher order of differencing to make it stationary. Next step is to find the order of differencing to convert the data from non-stationary to stationary. I performed this using the most widely used Augmented Dickey-Fuller Test by setting the critical value p = 0.05 and Null Hypothesis as “Data is not Stationary”. If the data is nonstationary , it must be differenced. If the ADF value was less than the critical value, then it means the data is stationary and I can accept the null hypothesis. In order to find the order of differencing, ndiffs() method was used to obtain the order of differencing. Both sales and traffic data were stationary with no seasonality or trend.

ARIMA model however predicted with very high RMSE.

**Model 3: Deep Learning – Recurrent Neural Network**

In the final method, Recurrent Neural Network was used for prediction. Here window slicing method was used. The previous 60 observations were used the predict the current sale/traffic and the window keeps sliding. The data was split into test and training sets and LSTM, Dense and Dropout layers were used. The data was flattened and then passed into model. The data was trained on 10 epochs and then predicted on test data. This model performed better than ARIMA model but still the prediction was off for most observations.