USER THROUGHPUT TIME SERIES FORECASTING  
 DATA SCIENCE PRACTICUM-2

Abhishek Gupta

agupta001@regis.edu

**User Throughput Time Series Forecasting**

**Problem/Situation**

Degradation in important cellular Network KPIs like User Downlink throughput can impact the satisfaction of end user negatively.

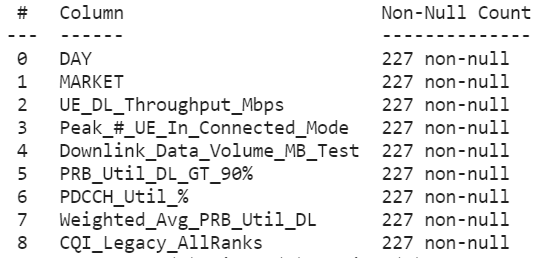
Devices have become more data centric rather than traditional calling or messaging services. Slow downloads speed can make a customer unhappy and hence it’s very important to provide customers with a good throughput speed. The operators try to maintain a minimum average throughput to make sure customer enjoy a good experience. It’s important to know in advance when the KPI might go below the targeted threshold so that operators can proactively work towards improving the Throughput by network optimization or adding more capacity and resources.

**Research Question**

Can time series forecasting help in predicting the Downlink Throughput KPI that can help cellular operators maintain the satisfaction level of customers by providing better experience?

**Data Collected**

The data was collected for 1 market for 227 days and which included Average User Downlink throughput along with other relevant KPIs. The Dataset has a total of 9 columns which are listed below. There were no null or missing values in the data.



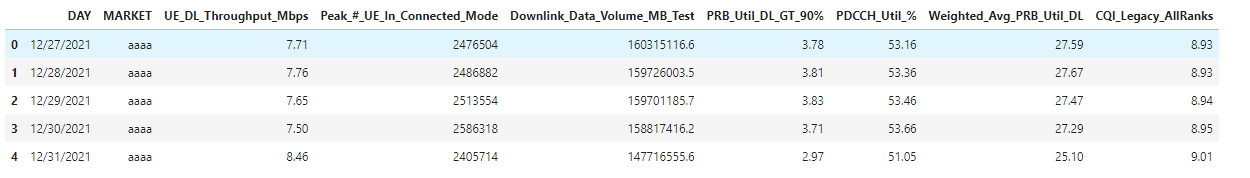
**Methodology**

To answer the research question presented above, I used Time Series Forecasting to predict the User Downlink Througput. For predicting the Throughput I used ARIMA, XGBoost and LSTM models.

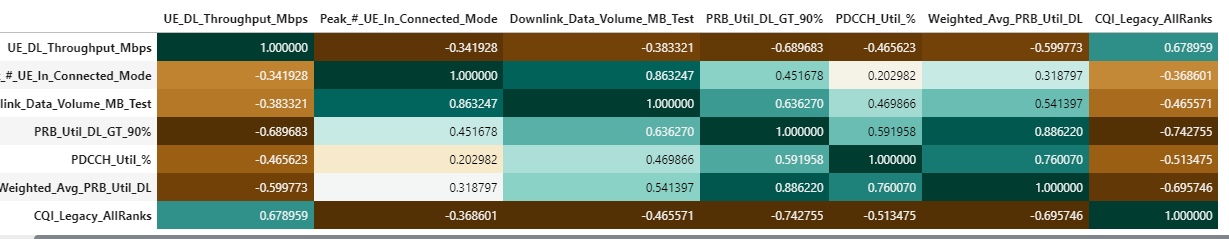
The performance observed with all the three models will be discussed below.

**Exploratory Data Analysis**

The data consists of 9 columns as shown below. The KPI to be predicted is “UE\_DL\_Throughput\_Mbps”

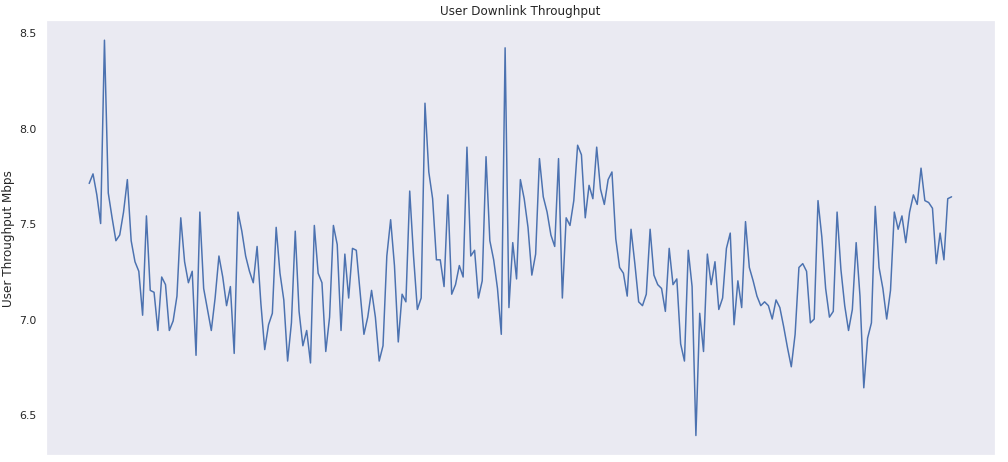


The below plot is the correlation plot and shows correlation between the KPIs. From the plot below we can see there is a good direct correlation of User throughput and CQI which means as CQI improves the throughput perceived by users improves. Also, the throughput KPI has high inverse correlation with PRB utilization greater than 90% and Average PRB Utilization. PRB are the resources given to user , so if the resource utilization is very high then user will get low throughput since users will not be able to scheduled faster because of resource limitation.



For this project, I will use univariate Time Series Forecasting and will predict Throughput without considering the impact of other KPIs.

The plot below shows the User Downlink Throughout for last 227 days.



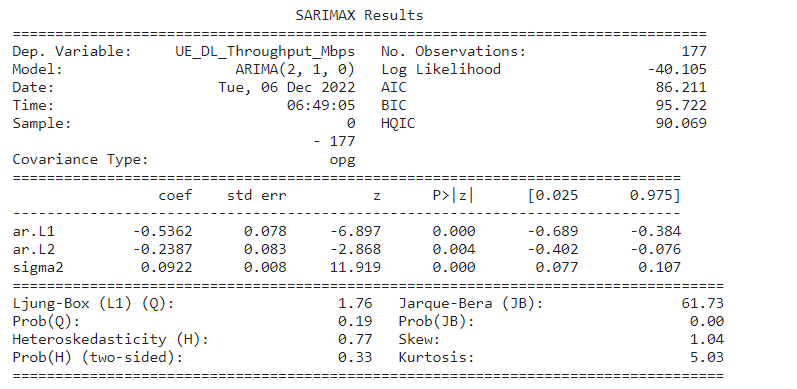
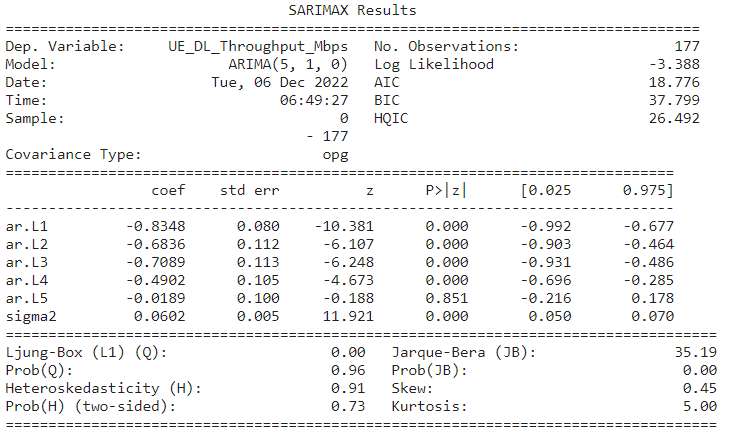
**ARIMA**

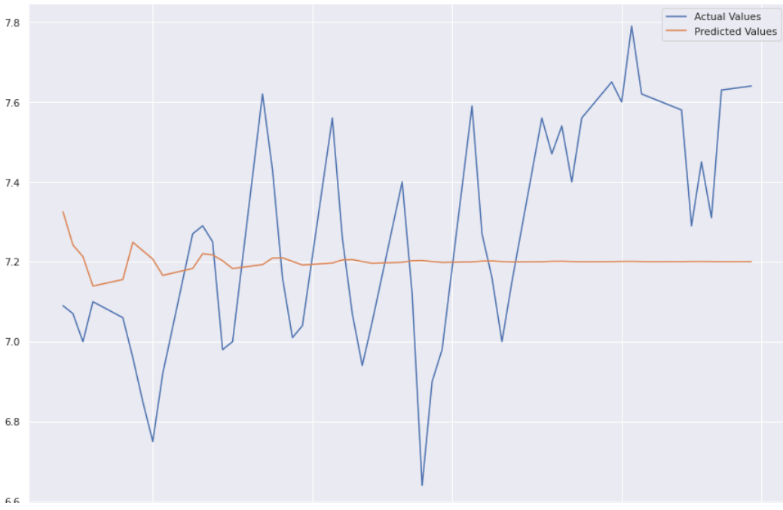
ARIMA stands for Auto-Regressive Integrated Moving Averages. ARIMA is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

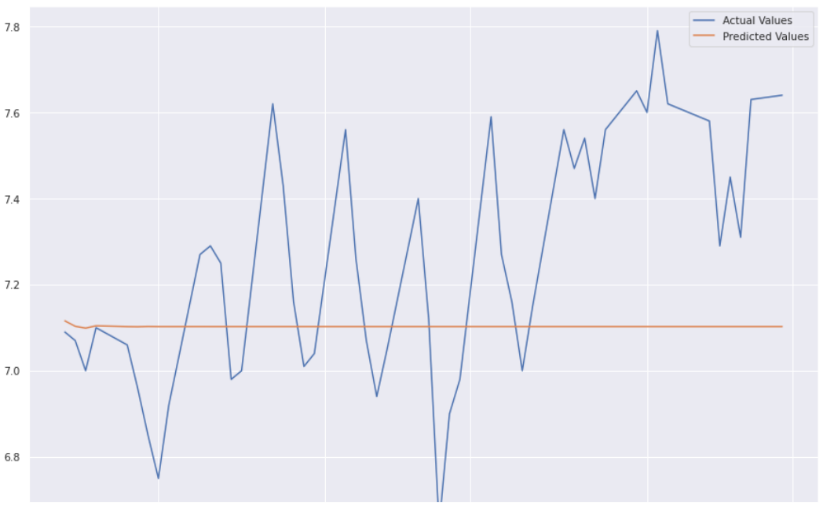
From the available 227 days of data, 177 days of data was used as training set and prediction was done on 50 days which was evaluated against the test data.

The rolling Statistics and Augmented Dickey-Fuller Test was performed to ensure that the timer series is stationary. The result showed that the time series we are using is stationary.

I used ARIMA(2,1,0) and ARIMA(5,1,0) to predict the Throughput values.





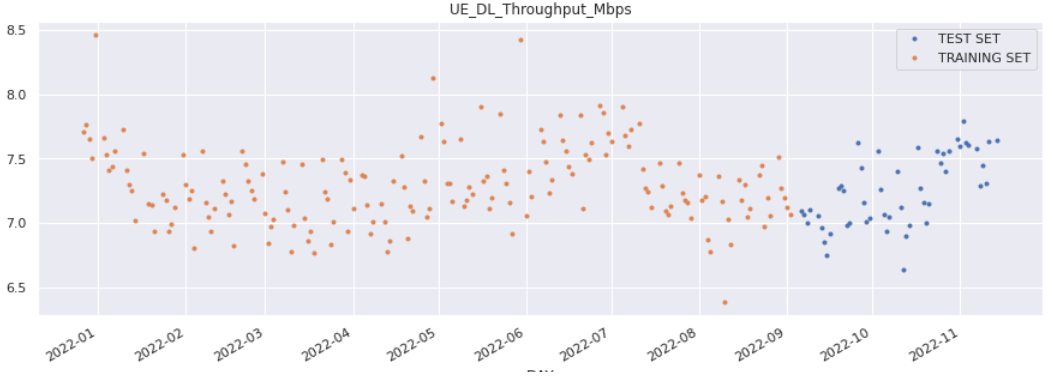


RMSE (Root of the Mean of the Square of Errors) observed for ARIMA(2,1,0) is 0.317 where the RMSE seen for ARIMA(5,1,0) is 0.267. The performance of ARIMA(5,1,0) seems to be slightly better than ARIMA(2,1,0).

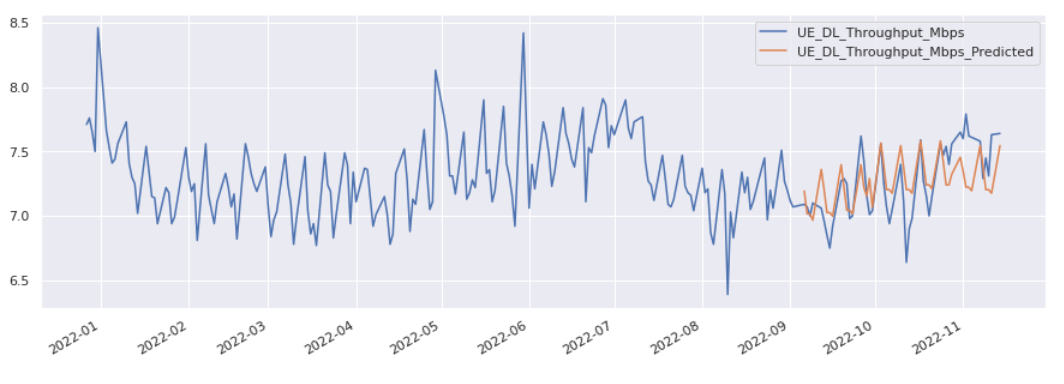
**XGBoost**

XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm. XGBoost is developed on the framework of gradient boosting.

Train/Test Split: The below graph shows the train and test split on the data. The training data set has 177 samples and test data has 50 sample points.



The below graph shows the original data and the predicted data in orange.

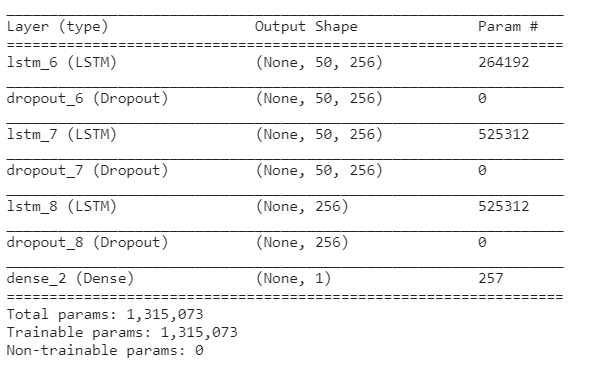


The RMSE observed for XGBoost is 0.224 which is better than ARIMA models.

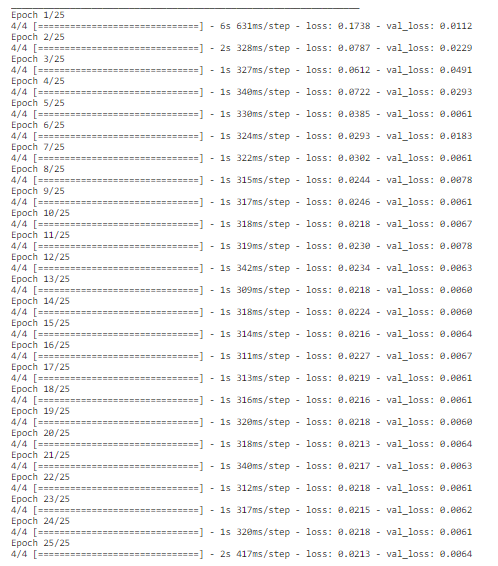
**LSTM**

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.

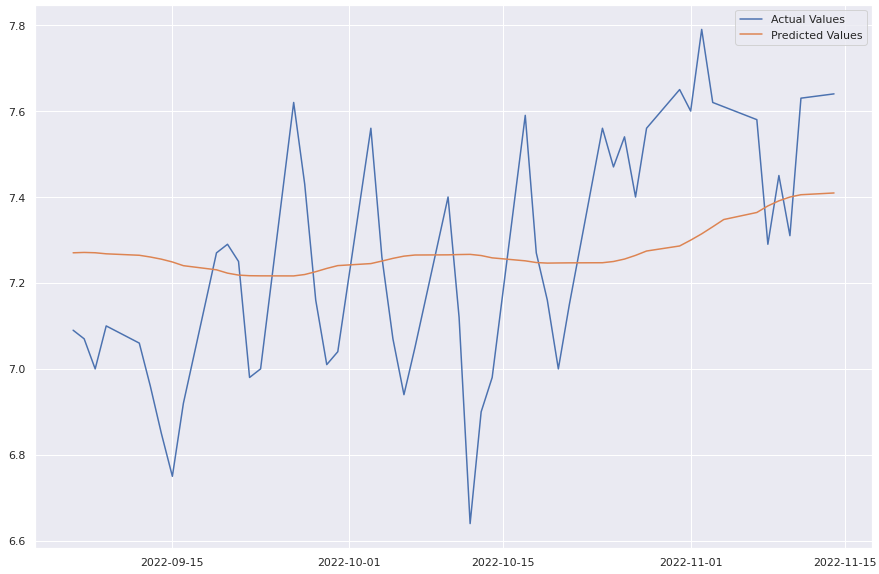
The dataset was split into train and test data. The training data was scaled between values of 0 and 1 using MinMaxScaler. Scaling the data is very important when working on neural networks, this model shortens the training time considerably.)



From the epoch below we can see that both the training loss and validation loss are reducing and finally becoming stable.



Below graph is for Actual Values vs Predicted Values



The RMSE observed with LSTM model was 0.263 which was better than ARIMA Models but slightly higher than XGBoost.

**Conclusion**

For this porject, I used Time Series Forecasting to predict UE\_DL\_Throughput in Mbps. I used 3 different models : ARIMA , XGBoost and LSTM to predict the Downlink UE Throughput and compared the performance of the 3 models.

For ARIMA, i used ARIMA(2,1,0) and ARIMA(5,1,0) and RMSE observed was 0.317 and 0.287 respectively. The RMSE observed was slightly better for ARIMA(5,1,0). Using XGBoost, the RMSE observed was 0.224 which was better than ARIMA models.With LSTM, the RMSE seen was 0.263 which was slightly worse than XGBoost but better than ARIMA models.

XGBoost performed better in predicting the UE\_DL\_Throughput followed by LSTM.

For future work, impact of other KPIs will be considered as well to predict the Downlink User througput and performance can be compared with the results we obtained from above methodology.