****

# **Data Science Candidates - Technical Assessment Test**

## **Problem Statement**

Mobile operators face congestion in their network use to factors like increased usage which lead to revenue loss for the company. We as SONALAKE data Scientist have to look into the cases of it (Data Analysis) and build ML model which forecast the usage in coming 14 days. Also, we need to check of congestion in those time and the cause which lead to it.

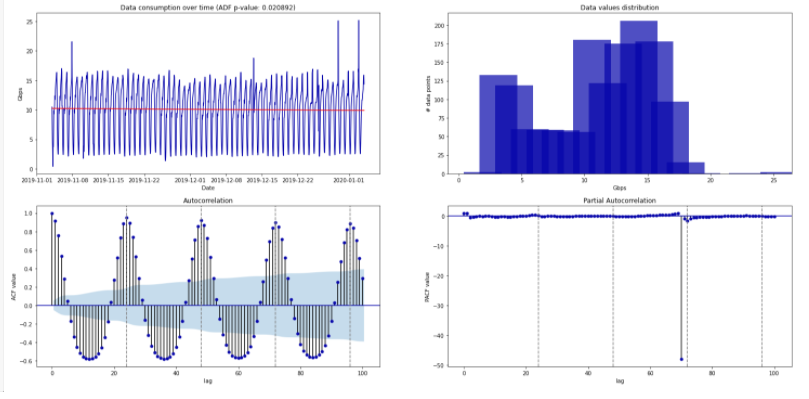
Any bandwidth greater than 16.40 Gbps is congestion.

The process is divided into few steps:

1. Data Analysis
2. Handling Missing value
3. Feature engineering
4. Model Creation
5. Forecasting the days bandwidth
6. Results
7. Summary and discussion

**Data Analysis:**

We will do some basic data Analysis to check for seasonality of data and the lag period in our data.



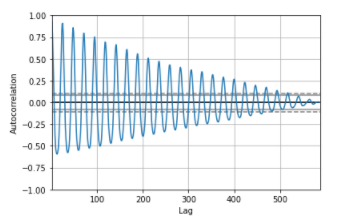
In the top left figure above we see the original behavior over time of the Data consumption. The p-value of the augmented Dick-Fuller (ADF) test, shown in the figure title, is a good indication that the time series is stationary.

As the three main criteria for a series to be stationary is satisfied here:

1. constant mean
2. constant variance
3. an autocovariance that does not depend on time.

The top right panel shows the value distribution of data consumption.

The bottom figures show the autocorrelation (ACF) and partial autocorrelation (PACF) functions. The ACF shows that the time series has a seasonal component at 24 hours (grey horizontal lines) with autocorrelation peaks which do not decrease over time, indicating the strength of this component. This is confirmed by the PACF.



As in correlation we compare two variable one against the other and check whether it is positive or negative correlated. Here auto correlation is correlation to itself i.e. how correlated is data from yesterday or someday.

From the figure above:

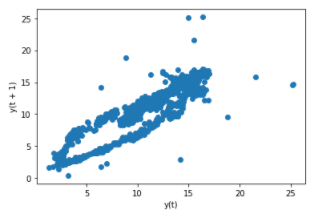
**------** 90% confidence interval; - - - - - 95% confidence interval

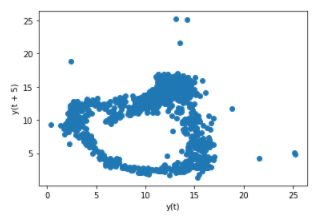
We can see their is correlation after every 24 hour, as the spikes are after 24 hour lag

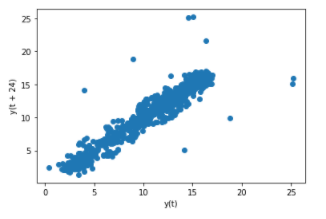
Shows that there is strong linearity or correlation between time frame above those interval

Gives an Idea for feature engineering on daily basis

Also, we will plot scatter plot with different lag values to confirm the same,

Lag value =1(Somehow linear correlated)

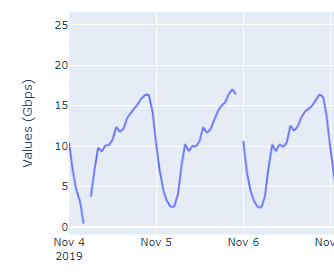
Lag value =5 (No correlation at all)

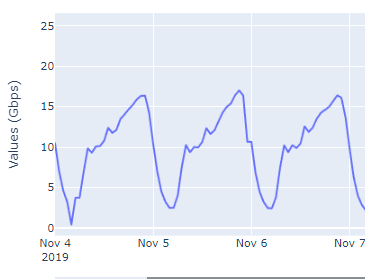
Lag Value=24 ( Linear Correlated)

**Handling Missing value**

We came to know that values have a linear correlation for lag 1(next/ pervious value) or lag 24( last day).

So for imputing the missing values we can go with ffill(forward fill), bfill(backward fill) or last day data filling. Imputing with mean values is not recommended as there is no pattern with other lag values.

 Missing value snapshot

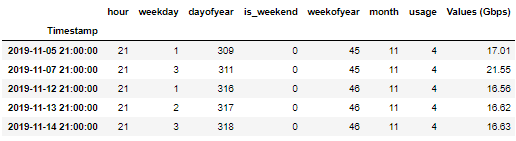
 After Imputing snapshot

**Feature engineering:**

New features are derived from the existing data which would be helpful in prediction.

We choose standard time series features like hour, weekday, month, etc.

* Usage as a feature is created by division of day into 4 quarters (1-4)
* Hour: hour in that day when data is recoreded
* Weekday: what is the weekday (0-Monday, 1- Tuesday, ……. 6 – Sunday)
* Dayofyear: day from starting of year
* Is\_weekend: 0/1 based on weekday or weekend
* Week of year: which week of the year
* Month: month when it is recored
* Usage: day into 4 quarters (1-4) i.e. 12 am to 6 am – 1, 6 am to 12 pm – 2, 12 pm to 6 pm -3. 6 pm to 12 am - 4



**Some Insights**

When data analysis is perform on this features:

* Congestion are seen on weekdays only on Tuesday and Thursday most, during the night only (usage=4)

**Training/testing data:**

The data is split into 90:10 ratio without shuffling the data with cross validation of 4.

**Model Creation:**

Once the data is split into training and test data, we will proceed with model creation using it, with Cross validation of 4.

We have worked on two model :

1. Ridge Regression ( base model)
2. XgBoost
3. **Ridge Regression:**

A basic model is created over to dataset using Ridge regression. The model is validated using standard cross validation technique adapted for time series data. The loss is checked through RMSE.

1. **Xgboost (e**X**treme Gradient Boosting):**

We choose Xgboost as it is an ensemble model which work in creating a strong model from weak models along with proper feature selection for modelling.

Now Xgboost model is created over the same data with hyper parameters optimization and it was expected to outperform. We used Bayesian optimization to find the optimal XgBoost algorithm hyper parameters as implemented by the hyperopt library. Cross-validation with rolling time window is used as metric for the optimization procedure. The optimized model is used to predict the values on the test set.

For Bayesian optimization, we are doing 10 maximum evaluation.

**14 days forecasting with ML models:**

I have used **Recursive strategy** for forecasting the 14 days ahead data i.e. for each forecasting step, this model(Xgboost) is used to predict one-step ahead and the value obtained from the forecasting is then fed into the same model to predict the following step till it complete 14 days.

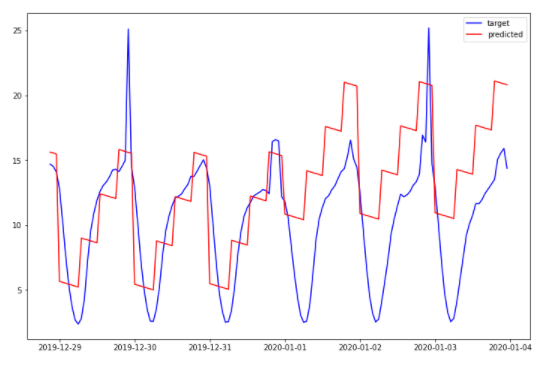
It is a cheap method to implement as same model is used again and again, which even sometime lead to forecasting error accumulation.

**Results:**

1. **Results for Ridge:**

Linear Regression RMSE cross-validation/test: 28.4484/4.5636

Using Linear Regression, yields a mean average percentage error (MAPE) of 28.44 from cross-validation and 4.56 when predicting on the much smaller test set.



1. **Results for Xgboost:**

The performance improved drastically using Xgboost Algorithm.

The best parameter for the modelling is

learning\_rate: 0.040230763312760744

n\_estimators: 300.0

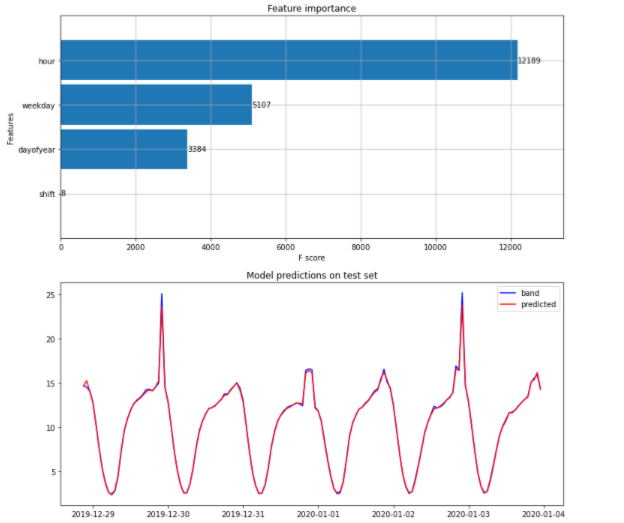
max\_depth: 4.0

sub\_sample: 0.8124340652406602

gamma: 85.0

The best model yields a mean average percentage error (MAPE) of 5.95 from cross-validation and 2.16 when predicting on the much smaller test set.

The predicted values on the test set and corresponding importance coefficients for the top features are shown below:



1. **14 days forecasting results:**

Using **Recursive strategy** for forecasting, we have plotted the prediction and the results of congestion are seen using the spikes whose values are greater than 16.4

Within these 14 days, saw 4 congestion which occurred at 10 pm ( as usual to the analysis done before).

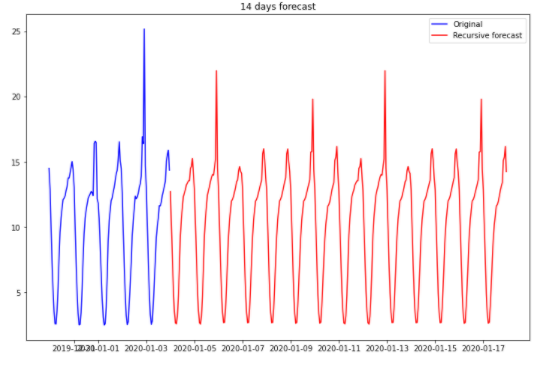
The 14 days forcasting showed 4 spikes(congestion) i.e.

2020-01-05 22:00:00 21.996199

2020-01-09 22:00:00 19.818935

2020-01-12 22:00:00 21.996199

2020-01-16 22:00:00 19.818935



**Conclusion**

* Xgboost model showed a very good improvement over Ridge Regression model with mean average percentage error (MAPE) of 5.95 compared to 28.44.
* Derived features which weighted most in modeling are hour, weekday, and day of year.
* The results showed that most congestion occur on weekdays only on Tuesday and Thursday manly.
* Congestion are seen mostly between 9 pm to 11 pm (usage=4)

**Solution:**

It is not preferred for the company to set up now site for this as the congestion is not occurring at daily basis and all the time. It is seen mostly on few weekdays and at night, probable reason may be people are free at night when the used their gargets most.

The solution can be to increase the threshold during the night so as to cover the bandwidth.