**Problem Statement & Objective:**

The goal for this dataset is to forecast the spatio-temporal traffic volume based on the historical traffic volume and other features in neighboring locations. Specifically, the traffic volume is measured every 15 minutes at 36 sensor locations along two major highways in Northern Virginia/Washington D.C. capital region. The 47 features include: 1) the historical sequence of traffic volume sensed during the 10 most recent sample points (10 features), 2) week day (7 features), 3) hour of day (24 features), 4) road direction (4 features), 5) number of lanes (1 feature), and 6) name of the road (1 feature). The goal is to predict the traffic volume 15 minutes into the future for all sensor locations. With a given road network, we know the spatial connectivity between sensor locations.

**Assumptions & Solution Design:**

1. First, we imported the dataset using the python scipy module capable of reading Matlab format files.
2. Then, we separated training and testing data from the imported dataset, finally converted them to a data frame object, and appended the quarter and spatial Information for smooth preprocessing and visualization.
3. Additionally, we have appended the number of connections for a spatial location derived from tra\_adj\_mat and created 47 features in that data frame that represents each feature described in the data description.
4. Target formulation - We have derived a feature called “next\_traffic\_flow” by getting the location-wise following traffic flow for the next 15 minutes.
5. Finally, we removed any nan entries from the data frame.
6. From basic statistics, Feature 0- Feature 9 is continuous features, and Feature 10- Feature 47 is a binary indicator variable except Feature 39, which has five unique values. These features represent the hours of the day. However, as this is a binary indicator feature, we have dropped them as the other 23 features are sufficient to represent the hour of the day.
7. We have plotted the histogram for each numeric feature in the dataset and concluded that most traffic flow is in the range of (0.0 to 0.1).
8. We have plotted Feature 0 and Next Traffic Flow and observed a primarily linear relationship.
9. We also observed that Next Traffic Flow is proportional to the number of connections for a particular spatial location.
10. Then, we performed correlation analysis on numeric features and observed that most of the features are very highly correlated. We take a threshold of 0.80 and only selected Feature 0 and Feature 8, current traffic\_flow as final numeric features, and all the categorical features except Feature 39 as final categorical features for baseline model building.
11. We have created a ridge regression pipeline using sklearn and found the best hyperparameter alpha using grid search.
12. Then, we created our baseline model using the optimized alpha and all the final features selected for model building.
13. Then, based on the model coefficient returned by the baseline model, we selected the top 10 most important features and finally created the final model using only 10 Features.
14. As we have used a regression-based model for predicting minimal traffic flow, the model can predict negative traffic flow as it does not respect the bounds of 0. To overcome this situation, we capped those negative values to zero.
15. Finally, We evaluated our model on test data and got the below-mentioned results-

Symmetric mean absolute percentage error -> 11.22

Root Mean Squared Error -> 0.0013

Mean Absolute Error -> 0.0252

r2\_score -> 0.9678

Adjusted r2\_score -> 0.9678

1. Plotted the residuals normally distributed with a mean of zero.
2. We also evaluated our model for each spatial location and plotted the Actual vs. Predicted following traffic flow.
3. Except for location 28, the model is working very well. Upon investigating, we observed location 28 having very volatile actual traffic flow. Our model could not predict those values correctly, resulting in a poor R-squared score.
4. Finally, we took the quarter hour = 840 information from the test dataset for all 36 spatial locations and forecasted the next 15 minutes of traffic flow using our model.

**Evaluation Metrics Used:**

* **Symmetric mean absolute percentage error** is an accuracy measure based on percentage (or relative) errors. It is usually define as follows:



* **Mean absolute percentage Error (MAPE)**, is a measure of prediction accuracy of a forecasting method. It usually expresses the accuracy as a ratio defined by the formula:



* **Root Mean Square Error (RMSE)** is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. The formula is:

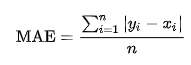
[rmse](https://www.statisticshowto.com/wp-content/uploads/2016/10/rmse.png)

Where:

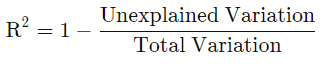
• f = forecasts (expected values or unknown results),

• o = observed values (known results).

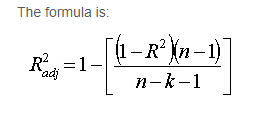
* **Mean absolute error (MAE)** is a measure of errors between paired observations expressing the same phenomenon. MAE is calculated as the sum of absolute errors divided by the sample size-



* **R-squared (R2)** is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.



**Adjusted R2** is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.



where:

N is the number of points in your data sample.

K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

**Conclusion:**

We evaluated our model on test data and got the below-mentioned results-

Symmetric mean absolute percentage error -> 11.22

Root Mean Squared Error -> 0.0013

Mean Absolute Error -> 0.0252

r2\_score -> 0.9678

Adjusted r2\_score -> 0.9678

Also we evaluated our model on all the spatial location individually and observed that Except for location 28, the model is working very well. Upon investigating, we observed location 28 having very volatile actual traffic flow. Our model could not predict those values correctly, resulting in a poor R-squared score.

**How can we improve it?**

• We can use non linear ensemble regression techniques like random forest or gradient boosting regression to improve the overall performance of the model.