Data Mining (DM) Assignment

Technical Design Document

Date: 27-08-2022

Authors: Group70

Prepared for: Data Mining (DM) Assignment 1

Document type: Technical Design Document (TDD)

Document Version: 1

This document is intended to help manage the delivery of an assignment and is provided for academic and technical review only. The information in this document is submitted for requirements discovery and submission purposes only.

# Contents

1. Problem statement 1
2. Data understanding and data preparation 2
3. Exploratory Data Analysis (EDA) 2
4. Significant Feature Ranking 8
5. Choosing machine learning algorithms for model training 9
6. Model training and evaluation 9
7. Model selection 10
8. Observation 10

## Identified and addressed the data quality issues 10

1. Problem statement

Traffic Flow Prediction Dataset: 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 the 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) weekday (7 features), 3) hour of the day (24 features), 4) road direction (4 features), 5) a number of lanes (1 feature), and 6) name of the road (1 feature). The goal is to predict the traffic volume for 15 minutes into the future for all sensor locations. With a given road network, we know the spatial connectivity between sensor locations.

## Data understanding and data preparation

This section details the data understanding and preparation of the traffic flow prediction. In this section, created the train and test dataframe from the traffic flow prediction dataset for machine learning. Restructured the dataset and added features like sensor\_id derived from the ‘36 sensors’ and merged the predictor and target/label variables. The following two functions are used to create the dataframe from the traffic flow prediction dataset.

1. fn\_create\_label()

This function is used to create the label/target variable from the traffic flow prediction dataset. The create\_label() function takes input as df\_name- Name of the Dataframe and y\_size- the size of contiguous quarter hours. This function returns the result as the df- Dataframe.

1. fn\_concat\_x()

fn\_concat\_x() function is used to create the predictor variable. The fn\_concat\_x() takes input as df\_name- Name of the Dataframe and y\_size: the size of contiguous quarter hours. The fn\_concat\_x() function returns the result as the df- Dataframe.

Performed the data understanding by printing a few rows of the dataframe and checked the missing values and data type etc. of the traffic flow dataset.

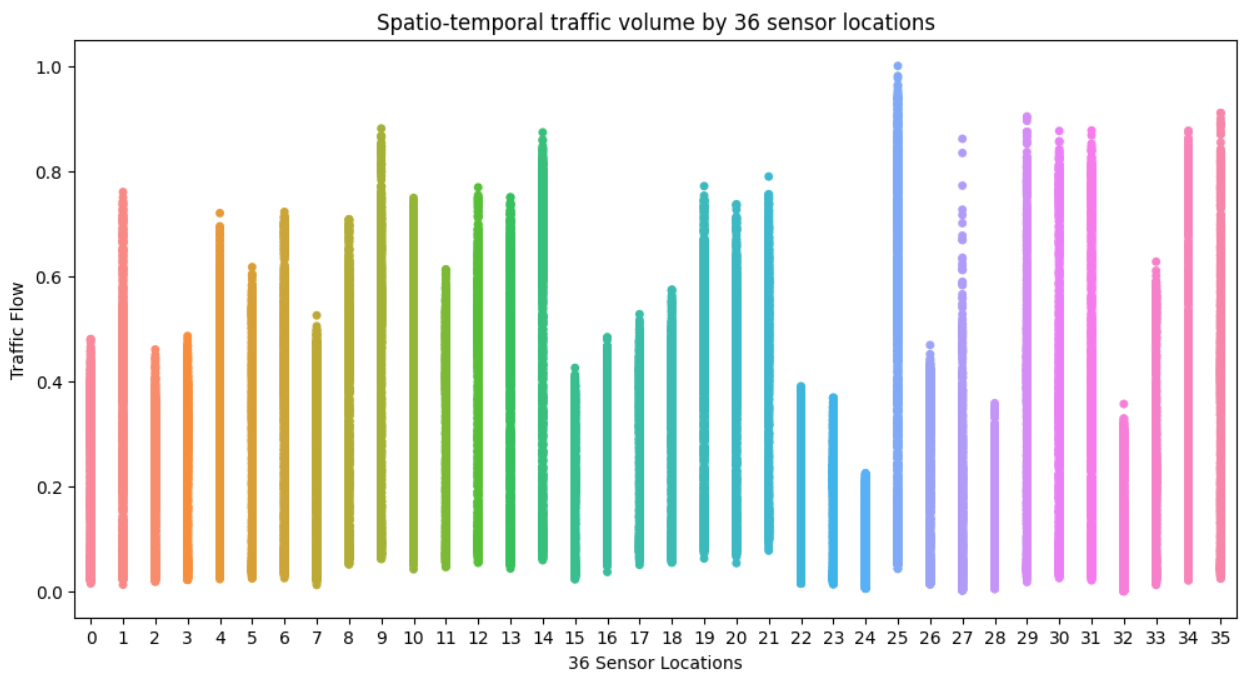
Export the merged training and testing dataset into CSV file format for further analysis and analytics.

## Exploratory Data Analysis (EDA)

This section discovers the trends and patterns in the dataset and understands the relationship between various variables.

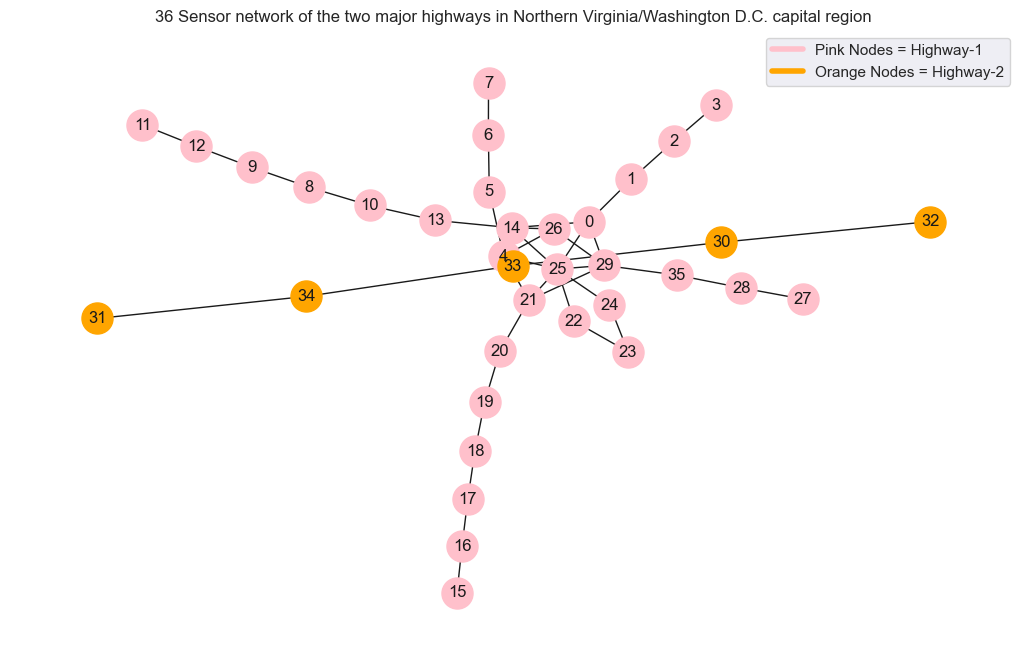
Exploratory Data Analysis (EDA) is an approach we used to analyze the data using visual techniques. EDA is used to check assumptions with the help of statistical summaries and graphical representations.

1. Spatio-temporal traffic volume by 36 sensor locations



The above plot shows the Spatio-temporal traffic volume by 36 sensor locations. Specifically as given, the traffic volume is measured every 15 minutes at 36 sensor locations. In the above plot, we are able to see that sensor\_id 23 has the lowest traffic flow and sensor\_id 25 has the highest traffic flow.

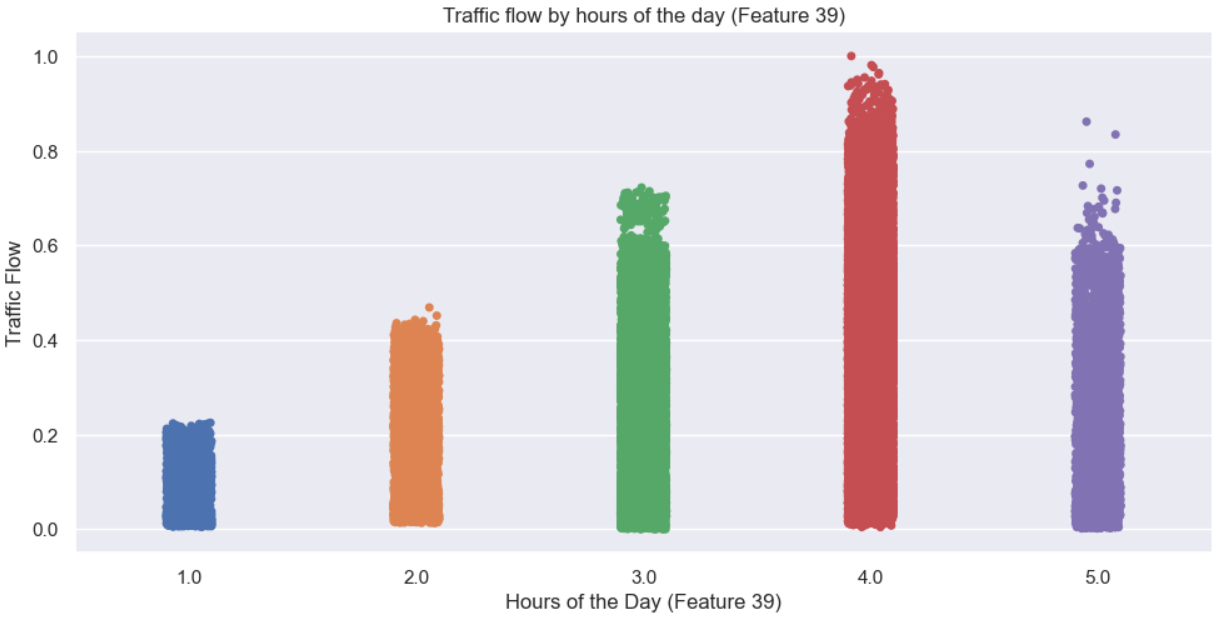
1. 36 Sensors network of the two major highways in Northern Virginia/Washington D.C. capital region



The Tra\_adj\_mat in the dataset is an adjacency matrix of size 36x36. The matrix represents an undirected graph with 36 nodes as 36 sensor locations on two major highways. Edges between the nodes represent a direct path that exists between two sensor locations through the highway. We performed an analysis on road names using Feature 47 of training data to separate sensors into highway 1 and highway 2 sensors.

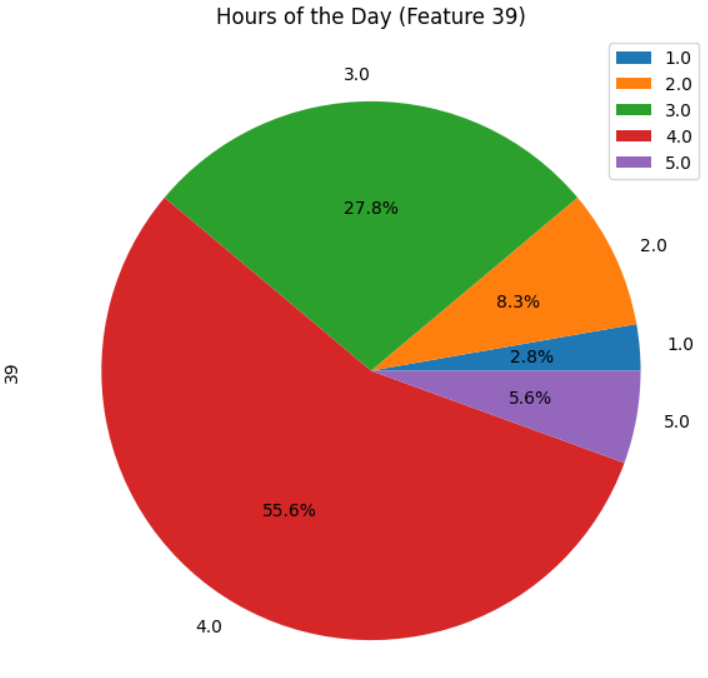
The above plot shows that Highway-1 (0.0 in Feature 47) is represented by pink nodes with respective node labels as 0, 1, 2, etc. and highway-2 is represented by orange nodes with respective node labels as 30, 31, 32 etc. From the above plot, it can be inferred that highway-1 has 31 sensors whereas highway-2 has 5 sensors.

1. Traffic flow by hours of the day (Feature 39)



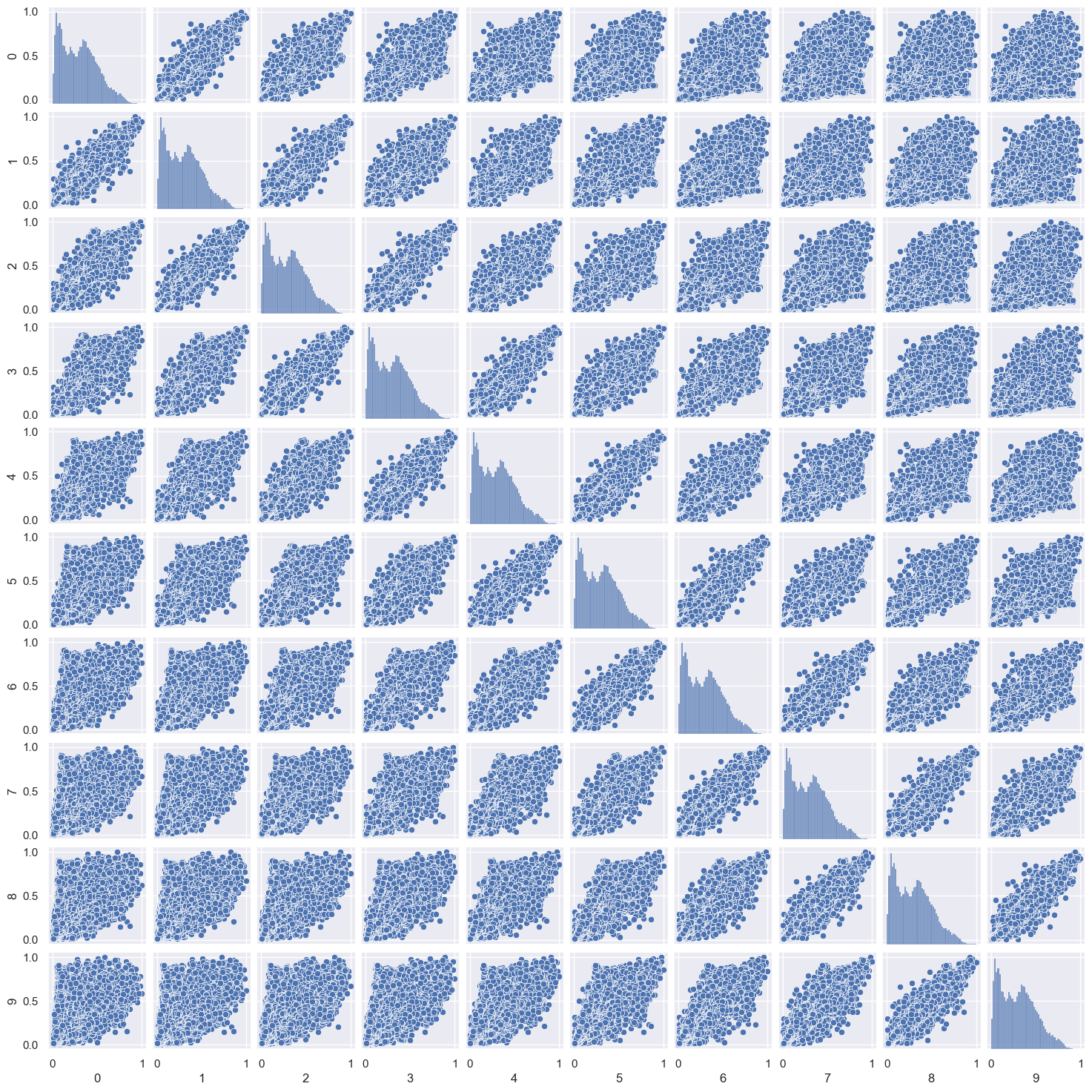
The above plot shows the traffic flow by hours of the day (Feature 39). In the given dataset the hour of the day has a total of 24 features and all the other features are in binary except Feature 39. The above plot presents that the traffic flow is highest at 4 hours of the day whereas the lowest is at 1 hour of the day.

1. Hours of the Day (Feature 39) in percentage



The above pie plot is divided into slices to illustrate the numerical proportion of the hours of the day (Feature 39). In the above plot, we can see the Four hours proportion is highest followed by Three hours compared to other hours of the day.

1. Pair plot of the historical sequence of traffic volume sensed during the 10 most recent sample points (10 features)



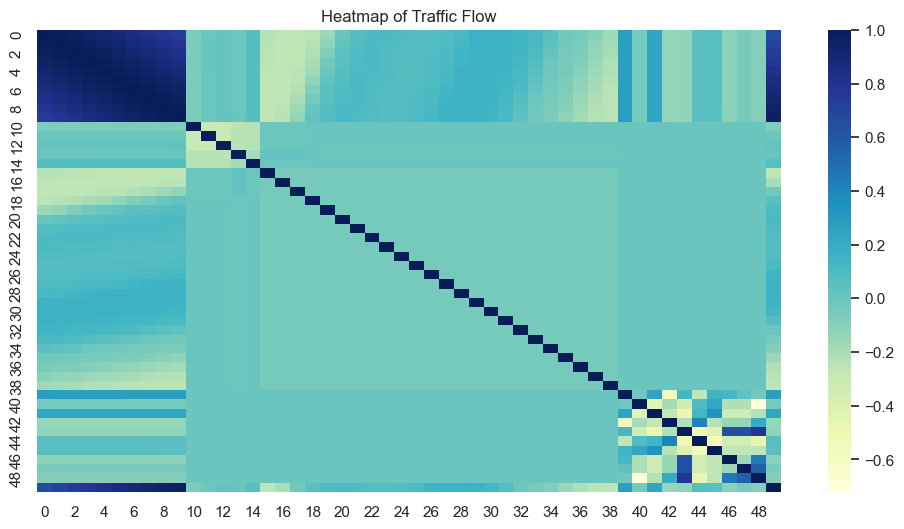
The above plot shows the correlation of the historical sequence of the traffic volume. In the above plot, we can see the positive correlation in the many features like feature 0 and feature 1, feature 8 and feature 9 and so on.

1. Linear regression model fit plot



The above plot shows the linear regression model fit plot. According to the given traffic flow prediction dataset, the predictor variable is Feature 9 and the target/label variable is traffic flow shows a positive [correlation](https://www.google.com/search?rlz=1C1YQLS_enIN907IN907&sxsrf=ALiCzsaVs9B5vTLB6kbcWyADqvjPxo6DVg:1661505730414&q=positive+relationship+or+correlation&spell=1&sa=X&ved=2ahUKEwjs1-mWl-T5AhXXR2wGHeuDBAkQkeECKAB6BAgBEDY).

1. Heatmap of Traffic Flow



The above plot shows the correlation of all features. We can see the highest correlation in features (feature\_0 - feature\_9).

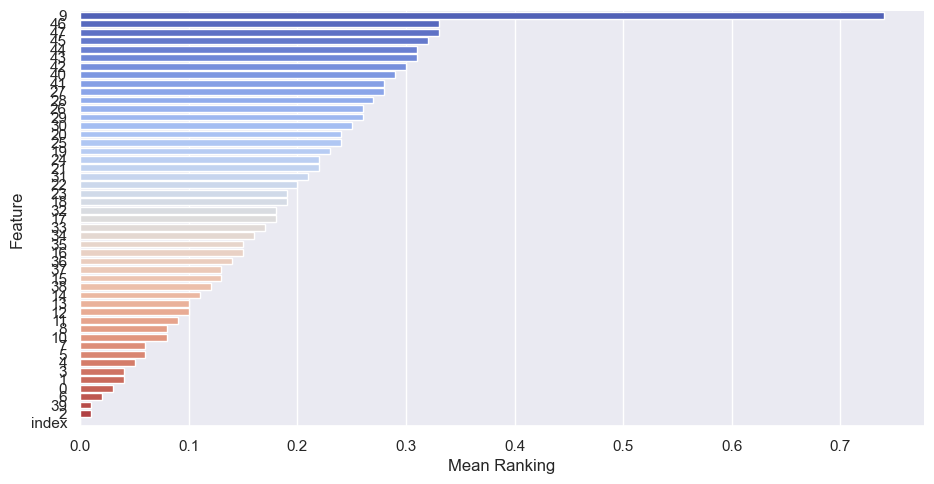
## Significant Feature Ranking

This section details the importance of features for robust machine learning model building. Feature selection is the method of reducing the input variable to the model by using only relevant data and getting rid of noise in the data.

Build the python function for features ranking name called feature\_ranking(). This function is used to select relevant features for the machine learning model. feature\_ranking() takes the input as X- Predictor variable, Y - Label/target variable and colnames - Name of the predictor variable. feature\_ranking() return the result as the meanplot - mean ranking of the predictor variable for features selection/ranking.

feature\_ranking() method can be used for feature selection, variable selection, attribute selection and also for variable subset selection. In another word, feature selection is the process of selecting a subset of relevant features for model building.

The following plot shows the mean ranking of features of the traffic flow prediction dataset



In the above plot, we can see that Feature 9 is most relevant for model building compared to the other features.

## Choosing machine learning algorithms for model training

This section details choosing the machine learning algorithms for model training. Choosing ML algorithms is one of the important steps in data mining. The choice of the machine learning algorithms is based on the given problem statement. For this linear problem statement chosen models are

1. K-Neighbors Regressor

2. Linear regression

3. Ridge regression models

4. Random Forest Regressor

5. Decision tree

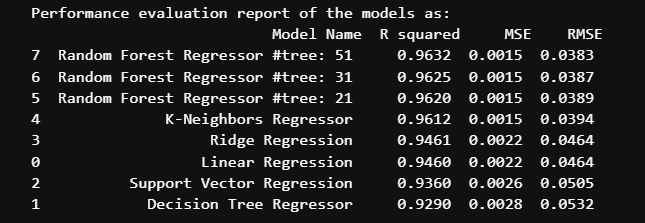
6. Support Vector Regressor

## Model training and evaluation

This section details machine learning model building and evaluation of the models. A prepared dataset was used to train and test the ML model performance. Split the training dataset into 70:30 ratios, where 70 portions were used for the model training and 30 portions were used for the model testing.

For model training, tried many permutations and combinations of features selection/ranking and trained and tested the chosen linear models. Finally, the Mean ranking of the features greater than equal to 0.05 (section 4) is selected for the ML model training.

The performance evaluation report includes the R2 Score, Mean Square Error (MSE) and Root Mean Square Error(RMSE) of the chosen ML models. The ML models performance evaluation report was sorted in descending order as:

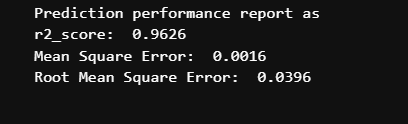


The above result shows that the Random Forest Regressor #tree: 51 performs well according to the performance/regression report as R2 Score, Mean Square Error (MSE), Root Mean Square Error (RMSE) compared to other models such as Random Forest Regressor #tree: 31, Random Forest Regressor #tree: 21, K-Neighbors Regressor, linear regression, Ridge regression models, decision tree, Extra Trees Regressor and Support Vector Regressor.

## Model selection and prediction

This section details the machine learning model selection and prediction of the given test traffic flow dataset.

According to the performance evaluation report, the Random Forest Regressor #tree: 51 model performance as the R2 Score, Mean Square Error (MSE) and Root Mean Square Error(RMSE) is 0.9632, 0.0015 and 0.0383 as shown in the above plot. K-Neighbors Regressor model is selected for further prediction and analytics.



## Observation

The overall prediction performance report of the Random Forest Regressor #tree: 51 models as R2 Score, Mean Square Error (MSE) and Root Mean Square Error(RMSE) is 0.9626, 0.0016 and 0.0396. Over time, with training, the model gets better at predicting.

## Identified and addressed the data quality issues

1. Noise on Day-of-Week Features (Features 10 to Features 16)

For many matrices of tra\_X\_tr, it was found that roughly 8.88% of 45396 records have 1 in two features out of 7 features (from features 1 to features 16) that represent the day of the week. As shown below, for Matrix 0 of tra\_X\_tr, both column 10 and column 15 include 1, which can be misleading as it can only be one day of the week (i.e., either Monday or Saturday).

Handled the Noise in day-of-week features in the model training and evaluation (section 6) by choosing the features that were most relevant for model building as discussed in Significant Feature Ranking (section 4).

2. Ambiguity around Feature 39 values in 24-hour\_of\_day features

The Ambiguity was found in the values of feature 39 which was represented as 24-hour\_of\_day features. This Ambiguity was handled in the model training and evaluation by dropping feature 39 from the model training and evaluation (section 6).

3. Date discrepancy between tra\_Y\_te and tra\_Y\_tr

* The date format is misleading as to whether it is YYYY-MM-DD or YYYY-DD-MM.
* tra\_Y\_te is for 840 quarter-hours which are roughly 8.75 days of data and tra\_Y\_tr is for 1261 quarter-hours which are roughly 13 days of data. The difference in starting dates between tra\_Y\_tr and tra\_Y\_te is roughly 30 days.

As per the problem statement to predict the Traffic flow, we choose the machine learning linear models for model training and evaluation and the Date discrepancy between tra\_Y\_te and tra\_Y\_tr was taken care of and handled during the model training and evaluation (section 6).