**Feature Engineering – Project Submission from Group-ID 25: 2019AIML083, 2019AIML066, 2019AIML016**

1. **Without building any model, what are the contributing factors for traffic slowness?**

Without building any model and by analyzing the record sets provided in the file, we understand that Hour (27 unique intervals) along with Broken\_Truck (having count of 70 non-zero records with 117 occurrences), Accident\_victim (having count of 43 non-zero records with 57 occurrences), Immobilized\_bus (having count of 35 non-zero records with 46 occurrences), Defect\_in\_the\_network\_of\_trolleybuses (having count of 17 non-zero records with 20 occurrences), Semaphore\_off (having count of 13 non-zero records with 17 occurrences), Lack\_of\_electricity (having count of 10 non-zero records with 16 occurrences), Point\_of\_flooding (having count of 6 non-zero records with 16 occurrences), Running\_over (having count of 14 non-zero records with 15 occurrences), Manifestations (having count of 7 non-zero records with 12 occurrences), Tree\_on\_the\_road (having count of 7 non-zero records with 12 occurrences) and Vehicle\_excess (having count of 6 non-zero records with 10 occurrences) are the contributing factors for the traffic slowness.

**Python script file or Jupyter notebook containing all the code for the proposed solution with proper comments**

**Activity 1: Data Understanding and Preparation along with EDA**

Approach: File Processing and Data Pre-Processing

Please refer file named # **FE\_Project.ipynb** for activities performed as part of Data understanding and Preparation along with Exploratory Data Analysis.

Step1: We have converted the given .arff file into .csv file format before doing the pre-processing work.

Step2: Data Quality Issues (Missing Values, Invalid Data, Outliers)

We have identified the number of missing values (*along with noise in data*) by comparing the number of records in each column and found missing values for Columns: Accident\_victim (2 missing value), Running\_over (1 missing value), Incident\_involving\_dangerous\_freight (2 missing value), Manifestations (1 missing value), Defect\_in\_the\_network\_of\_trolleybuses (2 missing value), Semaphore\_off (1 missing value).

Kindly note, we have passed ["-1", " ", ""] as missing value in the code by considering these as list of values and then replaced such values appropriately.

Post identification of missing values, we have removed spaces (wherever applicable) from the file.

Step3: Evaluating the non-contributing features/attributes based the Column values before applying PCA

Before implementing Principal Component Analysis, we have removed/dropped the following 5 Columns, considering they are non-contributing / non-influencing factors. Please see below details against each column reasoning for removal:

1. Fire\_vehicles: This is dropped as it has only 1 occurrence with Yes (i.e. 1), so not a contributing attribute/feature.
2. Occurrence\_involving\_freight: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
3. Incident\_involving\_dangerous\_freight: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
4. Fire: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
5. Intermittent\_Semaphore: This also has only 1 occurrence having 1, so not a contributing attribute/feature.

Now, we are left with 12 Columns on which we will apply PCA to further reduce dimensions/features towards model building activity.

Step4: As part of data pre-processing work, we have converted the categorical value to numerical value i.e. for the Column 'Vehicle\_excess', we have replaced 'T' with 1, 't' with 1, 'F' with 0 and 'f' with 0

Similarly, we have considered every 30 minutes slots starting from 7:00 as interval 1, 7:30 as interval 2, 8:00 as interval 3 and so on and converted the Column ‘Hour’ value accordingly in the code as numeric value, so we *have 27 interval* from ‘Hour’ perspective for the given 5 days from Monday to Friday between December 14, 2009 to December 18, 2009.

Note: The column ‘Fire\_vehicles’ is not required to be converted as we have already dropped this column.

We have taken an approach of finding a suitable slot/interval and hence Day is not considered as exclusive parameter in the computation. We have built a model to predict slowness in traffic for interval / hour and this will be applicable for all the Days.

Step5: Impute missing values: We have identified the missing value counts (wherever applicable) and imputed these missing values with appropriate values as under:

1. Accident\_victim (2 missing value): Has missing values for 12:00 and 19:30 Hours respectively. These values have been replaced with floor of mean.
2. Running\_over (1 missing value): Has been replaced with 0.
3. Vehicle\_excess: 1 value is replaced with 0.
4. Manifestations (1 missing value): Has one negative value (-1) at 13:00 Hour. This has been replaced with 0.
5. Defect\_in\_the\_network\_of\_trolleybuses (2 missing value): It has 2 blank values at 9:00 and 18:30 Hour respectively along with one negative value (-8) at timestamp 7:00 Hour. These values are replaced with 0.
6. Semaphore\_off (1 missing value): Has one missing value at 19:00 Hour. This is also replaced with 0.
7. Note: For columns ‘Incident\_involving\_dangerous\_freight’ & ‘Fire\_vehicles’, missing value imputation are not required as these columns were already dropped.

Step6: We have removed duplicates as part of next step.

Step7: Check for Outliers: In order to scale the given data, we have checked for outliers by using Boxplot visualization in the code for relevant columns and then removed those outlier values. Such as, we have identified the outlier value at timestamp 12:30 for Broken\_Truck and at timestamp 15:00 for Point\_of\_flooding and removed these rows.

Broken\_Truck: Contains outlier value as 1000 for 12:30 Hour and removed the entire row corresponding to this outlier value.

Point\_of\_flooding: Contains outlier value as 2000 for 15:00 Hour and removed the entire row corresponding to this outlier value.

Now, we are left with 132 record sets along with 13 Columns.

Identification of the factors causing the traffic slowness

Step 8: Feature Selection: In order to perform feature selection, we have considered ‘Slowness\_in\_traffic\_percent’ as response variable y, so we have 12 columns.

We have Identified features (Hour not included) that have high co-relation with respect to target variable (Slowness\_in\_traffic\_percent) using heatmap and correlation matrix.

Bivariate approach:

Check co-linearity among the features and extract features that are highly corelated to 'Slowness in traffic', taking threshold as 0.1

With the above heatmap & correlation matrix analysis, we observed that the 9 major factors contributing towards the slowness of traffic are Tree\_on\_the\_road', 'Incident\_involving\_dangerous\_freight', 'Point\_of\_flooding', 'Defect\_in\_the\_network\_of\_trolleybuses', 'Semaphore\_off', 'Accident\_victim', 'Manifestations', 'Lack\_of\_electricity', 'Broken\_Truck'.

Multivariate approach:

Identify collinearity among the features selected from correlation matrix.

**Activity 2: Model Building and Evaluation**

Step9: Split the data into train and test model as 70:30 ratio.

Step 10: Now, we have normalized the features and applied PCA to convert it into 6 new components, then printed the PCA transformed data.

Afterwards, we have created the regression model as:

Coefficient= Theta1 = [-2.97350899 0.03327026 0.83567985 -0.30192371 0.3843474 -0.49133105]

Intercept = Theta0 = 51.85786465016669

Thus, we have developed Model 1 as under:

Y = 51.86 – 2.97X1 + 0.03X2 + 0.83X3 - 0.30X4 + 0.38X5 -0.49X6

And for this Model 1, we have printed the ‘Predicted Slowness Values’ along with ‘Score’, ‘Mean Squared Error’ and ‘Variance Score’.

MSE = 5.14

R square value = 0.70

Model 1 Score: 0.6966

**Note: As part of building Model 2, we will first identify the contributing factors towards slowness in traffic using RFE technique and then will build Model 2 with the identified features.**

1. **Are you able to confirm the above findings using any two modelling techniques? Give appropriate explanation for the same.**

We have developed Model 1 by applying PCA (Principal Component Analysis) on the 12 attributes: Hour, Immobilized\_bus, Broken\_Truck, Vehicle\_excess, Accident\_victim, Lack\_of\_electricity, Point\_of\_flooding, Manifestations, Defect\_in\_the\_network\_of\_trolleybuses, Tree\_on\_the\_road, Running\_over & Semaphore\_off , and transforming these attributes to 6 components.

With this we achieved a regression model Y = 51.86 – 2.97X1 + 0.03X2 + 0.83X3 - 0.30X4 + 0.38X5 -0.49X6 , having mean squared error as 5.14 with R square value as 0.70

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. Thus, we see that the 6 PCA components developed using 12 attributes is a 70% accurate model for prediction of slowness in traffic.

Here after, we move forward to apply Recursive Feature Elimination technique.

**Activity 3: Using RFE technique to identify the factors contributing the slowness in the traffic**

Principal Component Analysis just returns the principal components which are actually combinations of features. It won’t tell us which features are most important. To address this issue, Recursive Feature Elimination technique is used.

Approach

1. Eliminate the recursive feature
2. Identify important features by recursively removing low ranked features
3. Initialize Recursive Feature Elimination model for these important features (7 in our case)
4. Transform data using RFE
5. Fit the data to model
6. Apply PCA to transform the data
7. Create regression model

The regression model developed using RFE technique with PCA transformation is as under:

Coefficient = Theta1 = [-2.93737116 0.64908777 0.69099998 0.08554439 0.41700897]

Intercept = Theta0 = 51.44901382943826

Y = 51.45 – 2.94X1 + 0.65X2 + 0.7X3 + 0.08X4 + 0.42X5

And for this Model 2, we have printed the ‘Predicted Slowness Values’ along with ‘Score’, ‘Mean Squared Error’ and ‘Variance Score’.

MSE = 5.17

R square value = 0.69

Model 1 Score: 0.6944

Thus, we conclude that even after reducing the features to 7 (using RFE), accuracy remains the same. But model 2 is simpler and faster.

**NOTE**: We have used PCA that facilitates with feature extraction technique to build Model 1 and RFE, i.e. feature elimination technique approach to build Model 2 for the feature subset selection.

1. **Using Recursive Feature Elimination technique, what are the contributing factors for traffic slowness?**

Using RFE technique, the 7 features: 'Hour', 'Immobilized\_bus', 'Lack\_of\_electricity', 'Point\_of\_flooding', 'Manifestations', 'Tree\_on\_the\_road' and 'Semaphore\_off' are identified as the contributing factors for slowness in traffic.