**PROGRESS REPORT**

**Group No**.: Group 7

**Student Names**: Parikshit Dumbhare and Urjasvit Sinha

**Solution Design:**

In this project, we would be generating a prediction model of spot prices of fuels that in total account for New York State. This data set consist of the average weekly spot fuel prices several key data attributes, i.e. New York Harbor conventional gasoline, ultra-low sulfur diesel as well as West Texas Intermediate (WTI) and Europe Brent crude oil. This data set also has a monthly wholesale price for jet fuel that is included at a 3-month lag. For uniformity, the data has been standardized prior for further analysis. This is a supervised machine learning prediction problem. In this project, several prediction models will be built using the outcome variables NYHC Gasoline and Ultra-Low Sulfur Diesel. The models will be tested for accuracy and performance and the best fit model will be tested on the dataset which will predict the outcome variable. Along with this, we will be predicting Jet Fuel Price using Time-Series forecasting technique.

*Predictors (Input variables) x* -

West Texas Intermediate (WTI) and Europe Brent crude oil.

*Outcome (Output Variables) y* -

New York Harbor conventional gasoline, ultra-low sulfur diesel, NYS Jet Fuel.

**Algorithm Selection:**

Algorithms to be used for predicting response variable:

1. ***Linear Regression:*** This modeling method tries to classify the association by fitting the observed data into a linear regression between two variables. One variable is considered to be an independent variable, and the other is considered to be a dependent/ outcome variable.
2. ***Multiple Linear Regression(MLR):*** This modeling is similar to linear regression but here the explanatory/ independent variables are two or more unlike the linear regression. Hence, named as multiple linear regression. We will be using MLR to compare it with a linear regression model.
3. ***Random Forest Regression:*** This modeling technique has capability of both regression and classification using multiple decision trees. This is done using a technique called Bootstrap Aggregation/ Bagging. We would be using Random Forest Regression to predict the outcome variable.
4. ***Time Series Forecasting:*** This is a technique used for prediction of events completed over a series of time. NYS Jet Fuel prices which are updated every month can be predicted using this technique.

**Implementation:**

1. Load the dataset (.csv file) and use summary() function to describe the dataset structure.
2. Check the dataset for garbage/null values and cleaning the dataset.
3. Finding the appropriate response and predictor variables.
4. Visualization of the predictor variables and checking for normal behavior.
5. Finding correlation between the response variables and predictor variables thereby doing variable selection for the prediction model.
6. Dividing the dataset into training, testing and validation (60%,20%,20% respectively).
7. Using multiple algorithms and training the predicting variables on the training dataset and finding the best model which predicts with highest accuracy.
8. Taking the best fit model and testing that on testing part of the dataset and checking for overfitting of the model.
9. Measuring the accuracy of the model musing a confusion matrix.
10. Lastly using that best fit model for predicting response variable using validation data set.
11. Using ROC and lift chart for visual comparison of the model performance.
12. Finding residual errors using MAE, MAPE, MPE, RMSE, R square values.