**USE CASE STUDY REPORT**

**Group No**.: Group 07

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

**Executive Summary:**

In this project, we have developed prediction models of spot prices of fuels in the state of New York. This prediction model will prove beneficial for predicting gasoline and diesel prices according to price surges in crude oil barrels. The goal of this project is to apply several regression methods along with a forecasting method on the dataset and get the best model which will predict the prices of these fuels with highest accuracy. The dataset for used for making this project is taken from the New York State Energy Research and Development Authority (NYSERDA) which publishes a Weekly Transportation Fuels Report of average spot fuel prices for New York Harbor conventional gasoline and ultra-low sulfur diesel as well as West Texas Intermediate (WTI) and Europe Brent crude oil.

As for data pre-processing the data set was checked for null values and new columns Month and Year were extracted from dates provided and added to the data set for using it for time series forecasting. Also, visualizations were created from the data set for checking out patterns and trends for further analysis.

We have used various modeling techniques which are some of the significant machine learning methods. Initially after visualizations, simple linear regression was used for modeling gasoline and diesel through two different models. Then by choosing two parameters for each model, multiple linear regression was performed. This was followed by random forest regression. Along with the regression methods, we used time-series forecasting to predict the jet fuel prices.

At last, all the built models were analyzed using performance parameters such as RSME, MAPE, Min-Max Accuracy and errors and their residuals were plotted to evaluate the predictions made. In addition to evaluating prediction accuracy, K-Fold Cross Validation technique was applied to the regression models.

We were quite successful in predicting the prices of conventional fuels using the crude oil prices and found out that they were highly correlated in determining the prices. We believe that other factors like rapid world economic growth (demand and supply) or responses to conventional energy supplies or climate change or continuing conflicts in the Middle-East should also be taken into consideration. We recommend using these factors to predict the prices and further increase the diameter of the prediction of these conventional fuels.

**I. Background and Introduction**

In this project, we have built prediction models which predict the prices of the spot fuels in the State of New York. This data set consists of the average weekly spot fuel prices like 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 project is a supervised machine learning prediction problem. In this project, several regression models were built using the outcome variables NYHC Gasoline and Ultra-Low Sulfur Diesel. The models were tested for accuracy and performance and the best fit model was tested on the dataset to predict the outcome variable. Along with this, we also predicted Jet Fuel Price using Time-Series forecasting technique.

# II. Data Exploration and Visualization

We have used a dataset which is hosted by the State of New York. This dataset consists of 6 columns named “Date”, “NY Conventional Gasoline Spot Price ($/gal)”, “NY Ultra-Low Sulfur Diesel Spot Price ($/gal)”, “WTI Crude Oil Spot Price ($/barrel)”, “Brent Crude Oil Spot Price ($/barrel)”. Dataset was then expanded with two more columns extracted from the “Date” column and were named “month”, “year”. Further 6th,7th and 8th columns from the dataset were removed as they were not required for the analysis.

For finding out patterns and trends between the columns of the dataset, we plotted scatterplot for gasoline and diesel.

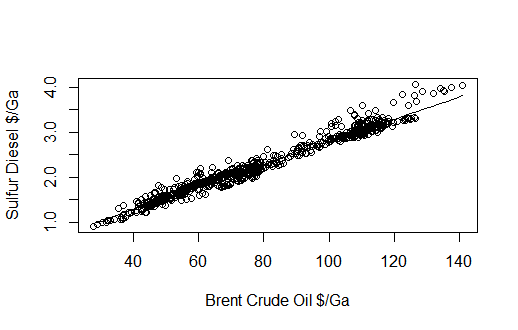


Fig. 1 Scatterplot (Brent Crude Oil vs Sulphur Diesel)

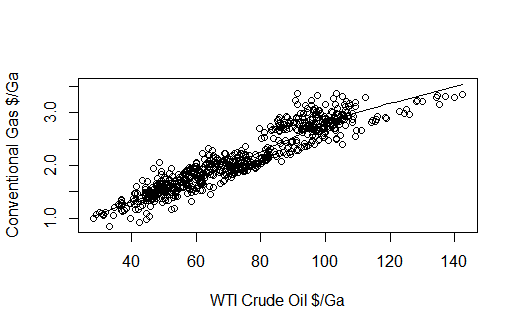


Fig. 2 Scatterplot (WTI Crude Oil vs Conventional Gasoline)

Then to check the normality of the response variables, density Plot was plotted. It was observed that response variables were normally distributed.

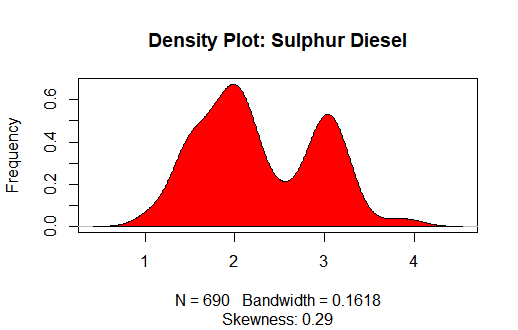


Fig. 3 Density plot (Sulphur Diesel)

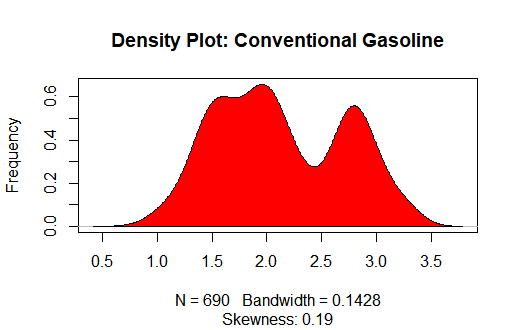


Fig. 4 Density plot (Conventional Gasoline)

We also plotted the correlation matrix for all five parameters to find out which attribute relates the most.

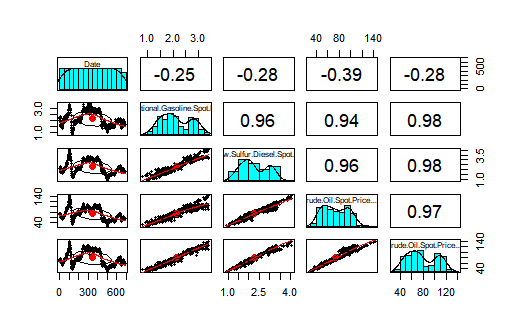


Fig. 5 Correlation Matrix

From the correlation plot, we inferred that the most important parameters such as gasoline and diesel are positively correlated to BRENT crude oil and WTI crude oil. It was also observed that all these parameters are negatively correlated to date.

# III. Data Preparation and Preprocessing

Data preparation and preprocessing are the most important factors of machine learning. So prior to applying modeling techniques, the dataset was checked for null values, which were 634 in total. And for applying data mining techniques, the dataset was divided into training, testing and validation sets. Also, for predicting the Jet Fuel Price using time series forecasting we selected Jet Fuel Price column from the dataset and by using the lubridate library we transformed the non-null values of the column into month, year and price.

# IV. Data Mining Techniques and Implementation

The problem can be solved by regression analysis. So here our two dependent variables Gasoline and diesel were predicted used two different crude oils i.e. WTI Crude oil and Brent Crude oil. The flow chart will describe the process of solving this problem:

Simple Linear Regression

Multiple Linear Regression

Random Forest Regression

Time – Series Analysis

RMSE

MAPE

MAE

K-Fold Cross Validation

Data Cleaning and Pre-processing

Sampling – Dividing Dataset into Training, Validation and Testing

Exploration and Visualizations

Data-Set

Performance Evaluation

Data Mining Techniques

Fig. 6 Flowchart of Project

# V. Performance Evaluation

To evaluate the performance of the data mining techniques and find the best one out of them, we firstly built the model using a training set and then the testing set was fitted to the model so as to check for the overfitting and accuracy of the models. Along with that several methods were used for performance evaluation such as Mean Absolute Error, Mean Percentage Error, Root Mean Square Error to measure the prediction accuracies and K-Fold Cross Validation to check for overfitting of the regression models (Linear Regression, Multiple Linear Regression and Random Forest Regression) on training and testing dataset. For further evaluation of the random forest model, we plotted error rate against the number of trees. As for time series forecasting, we tried to plot the residuals of the forecasted time series against the time for which the forecast was made.

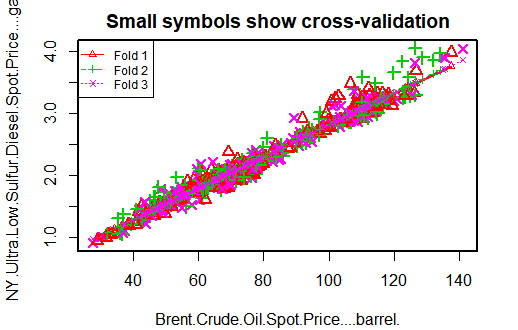


Fig. 7 Cross-validation plot for Sulfur Diesel

3-fold cross-validation plot between the sulfur diesel spot price and Brent crude oil price.

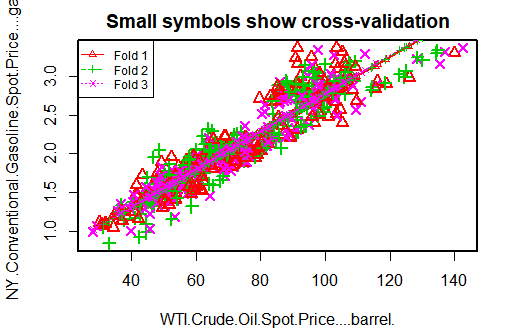


Fig. 8 Cross-validation plot for gasoline

3-fold cross-validation plot between the conventional gasoline spot price and WTI crude oil price.

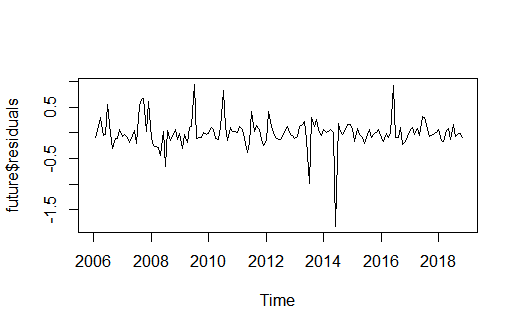


Fig. 8 Residual plot (time-series)

Time series future forecast residual plot. This residual plot shows residuals of jet fuel vs time.

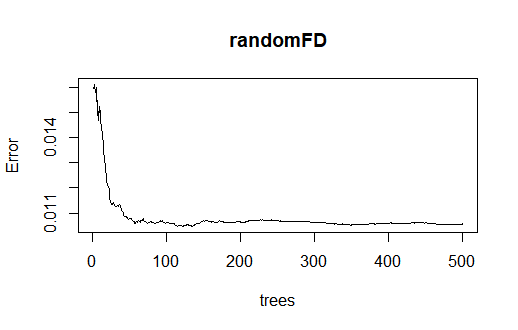


Fig. 9 Error vs the number of trees plot for sulfur diesel spot price in a random forest.

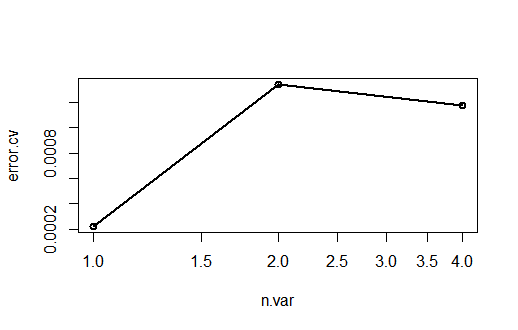


Fig. 10 Cross-validation for the number of variables and error rates.

# VI. Discussion and Recommendation

The study was quite successful after prediction of prices for conventional fuels such as Gasoline and Diesel. Various regression techniques such as Simple linear regression, multiple linear regression and random forest regression was used. It was observed that Multiple linear regression was better modeling technique as r-square value and better prediction accuracy. As we all know, Random forest regression needs big data set for better prediction, so on small dataset as used in this case study was not enough.

Hence, using this case study, predicting the prices of fuels used through prices of crude oil will play big role in economic development of New York state. The factors used in the study is very well evaluated and prediction models are cross-validated.

# VII. Summary

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. Using that best fit model for predicting response variable using validation data set.
10. Finding residual errors using MAE, MAPE, MPE, RMSE, R square values.

# Appendix: R Code for use case study

library(randomForest)

library(psych)

library(MLmetrics)

library(rsample)

library(forecast)

library(corrplot)

library(EnvStats)

library(timeSeries)

library(e1071)

library(DAAG) #Cross Validation(Linear Regression)

library(TTR)

library(caret)

library(ggplot2)

library(tidyverse)

library(lubridate)

library(tidyr)

library(dplyr)

library(ROCR)

library(party)

Transportation\_df <- read.csv("transportation-fuels-spot-prices-beginning-2006.csv")

head(Transportation\_df,5)

str(Transportation\_df)

summary(Transportation\_df)

sum(is.na(Transportation\_df$NY.Jet.Fuel.Price....gal.))

Transportation\_df$month <- month(Transportation\_df$Date)

Transportation\_df$year <- year(Transportation\_df$Date)

temp <- Transportation\_df %>%

dplyr::select(month,year, NY.Conventional.Gasoline.Spot.Price....gal.) %>%

na.omit() %>%

dplyr::arrange(desc(-year,month))

Transportation\_df <- Transportation\_df[,-c(6,7,8)]

##### GRAPHICAL ANALYSIS(Scatterplot, Density Plot, Correlation Plot) #####

#Scatterplots

scatter.smooth(x=Transportation\_df$WTI.Crude.Oil.Spot.Price....barrel.,

y=Transportation\_df$NY.Conventional.Gasoline.Spot.Price....gal.

,xlab = "WTI Crude Oil $/Ga",ylab = "Conventional Gas $/Ga")

scatter.smooth(x=Transportation\_df$Brent.Crude.Oil.Spot.Price....barrel.,

y=Transportation\_df$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

,xlab = "Brent Crude Oil $/Ga",ylab = "Sulfur Diesel $/Ga")

#Density Plot to Check Normality of Response Variables

plot(density(Transportation\_df$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.),

main="Density Plot: Sulphur Diesel", ylab="Frequency",

sub=paste("Skewness:", round(e1071::skewness(Transportation\_df$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.), 2))) # density plot for 'speed'

polygon(density(Transportation\_df$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.), col="red")

plot(density(Transportation\_df$NY.Conventional.Gasoline.Spot.Price....gal.),

main="Density Plot: Conventional Gasoline", ylab="Frequency",

sub=paste("Skewness:", round(e1071::skewness(Transportation\_df$NY.Conventional.Gasoline.Spot.Price....gal.), 2))) # density plot for 'speed'

polygon(density(Transportation\_df$NY.Conventional.Gasoline.Spot.Price....gal.), col="red")

#Checking for Correlation Between the Columns and Plotting them

pairs.panels(Transportation\_df)

############# RANDOM SAMPLING OF DATASET ##############

#Setting Seed and Dividing the Data Set into Training, Testing, Validation Samples

set.seed(123)

sample.train <- sample(seq\_len(nrow(Transportation\_df)),size = floor(0.60\*nrow(Transportation\_df)))

sample.test <- sample(seq\_len(nrow(Transportation\_df)),size = floor(0.20\*nrow(Transportation\_df)))

sample.validate <- sample(seq\_len(nrow(Transportation\_df)),size = floor(0.20\*nrow(Transportation\_df)))

trans\_train <- Transportation\_df[sample.train,]

trans\_validate <- Transportation\_df[sample.validate,]

trans\_test <- Transportation\_df[sample.test,]

#################### LINEAR REGRESSION ######################

linearC <- lm(NY.Conventional.Gasoline.Spot.Price....gal.

~WTI.Crude.Oil.Spot.Price....barrel.,data = Transportation\_df)

summary(linearC)

linearG <- lm(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

~Brent.Crude.Oil.Spot.Price....barrel.,data = Transportation\_df)

summary(linearG)

#Model 1

lm1<-lm(NY.Conventional.Gasoline.Spot.Price....gal.

~WTI.Crude.Oil.Spot.Price....barrel.,data = trans\_train)

plot(x=trans\_train$WTI.Crude.Oil.Spot.Price....barrel.,y=trans\_train$NY.Conventional.Gasoline.Spot.Price....gal.

,xlab = "WTI Crude Oil $/Ga",ylab = "Conventional Gas $/Ga")

abline(lm1,col="green")

#Summarizing

Model 1

summary(lm1)

AIC(lm1)

#Model 2

lm2<-lm(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

~Brent.Crude.Oil.Spot.Price....barrel.,data = trans\_train)

plot(x=trans\_train$Brent.Crude.Oil.Spot.Price....barrel.,

y=trans\_train$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

,xlab="Brent Crude Oil $/Ba",ylab="Ultra Low Sulfur Diesel $/Ga")

abline(lm2, col="red")

#Summarizing Model 2

summary(lm2)

AIC(lm2)

lt1 <- predict(lm1,data=trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.)

lt2 <- predict(lm2,data=trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.)

actual\_pred1 <- as.data.frame(cbind(actual=trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.,

predicted= lt1))

cor\_accuracy1 <- cor(actual\_pred1)

head(actual\_pred1)

actual\_pred2 <- as.data.frame(cbind(actual=trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.,

predicted= lt2))

cor\_accuracy2 <- cor(actual\_pred2)

head(actual\_pred2)

min\_max\_accuracy1 <-mean(apply(actual\_pred1, 1, min)/apply(actual\_pred1, 1, max))

min\_max\_accuracy2 <-mean(apply(actual\_pred2, 1, min)/apply(actual\_pred2, 1, max))

rmserr1 <- caret::RMSE(actual\_pred1$predicted,actual\_pred1$actual)

rmserr2 <- caret::RMSE(actual\_pred2$predicted,actual\_pred2$actual)

mape1 <- mean(abs(actual\_pred1$predicted - actual\_pred1$actual)/actual\_pred1$actual)

mape1

mape2 <- mean(abs(actual\_pred2$predicted - actual\_pred2$actual)/actual\_pred2$actual)

mape2

cvResults1 <- CVlm(data = Transportation\_df,form.lm = formula(NY.Conventional.Gasoline.Spot.Price....gal.~WTI.Crude.Oil.Spot.Price....barrel.),

m = 3,dots = FALSE,seed = 29,plotit = c("Observed","Redidual"),

main ="Small symbols show cross-validation",legend.pos = "topleft",

printit = TRUE)

attr(cvResults1, 'ms')

cvResults2 <- CVlm(data = Transportation\_df,form.lm = formula(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.~Brent.Crude.Oil.Spot.Price....barrel.),

m = 3,dots = FALSE,seed = 29,plotit = c("Observed","Redidual"),

main ="Small symbols show cross-validation",legend.pos = "topleft",

printit = TRUE)

attr(cvResults2, 'ms')

################## MULTIPLE LINEAR REGRESSION #####################

#Building Model on the Whole Data Set

Multi\_linearC <- lm(NY.Conventional.Gasoline.Spot.Price....gal.

~WTI.Crude.Oil.Spot.Price....barrel. + Brent.Crude.Oil.Spot.Price....barrel. ,data = Transportation\_df)

summary(Multi\_linearC)

Multi\_linearG <- lm(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

~WTI.Crude.Oil.Spot.Price....barrel. + Brent.Crude.Oil.Spot.Price....barrel.,data = Transportation\_df)

summary(Multi\_linearG)

#Building the MLR Models

#Model 1

Mlr1<-lm(NY.Conventional.Gasoline.Spot.Price....gal.

~WTI.Crude.Oil.Spot.Price....barrel. + Brent.Crude.Oil.Spot.Price....barrel.,data = trans\_train)

#Summarizing Model 1

summary(Mlr1)

AIC(Mlr1)

#Model 2

Mlr2<-lm(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.

~Brent.Crude.Oil.Spot.Price....barrel. + WTI.Crude.Oil.Spot.Price....barrel.,data = trans\_train)

#Summarizing Model 2

summary(Mlr2)

AIC(Mlr2)

#Predicting the Linear Models

Mlt1 <- predict(Mlr1,data=trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.)

Mlt2 <- predict(Mlr2,data=trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.)

#Calculating prediction accuracy

Multi\_actual\_pred1 <- as.data.frame(cbind(actual=trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.,

predicted= Mlt1))

cor\_accuracy1 <- cor(Multi\_actual\_pred1)

head(Multi\_actual\_pred1)

Multi\_actual\_pred2 <- as.data.frame(cbind(actual=trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.,

predicted= Mlt2))

cor\_accuracy2 <- cor(Multi\_actual\_pred2)

head(Multi\_actual\_pred2)

#Min Max Accuracy(Higher the better)

Multi\_min\_max\_accuracy1 <-mean(apply(Multi\_actual\_pred1, 1, min)/apply(Multi\_actual\_pred1, 1, max))

Multi\_min\_max\_accuracy2 <-mean(apply(Multi\_actual\_pred2, 1, min)/apply(Multi\_actual\_pred2, 1, max))

#Root Mean Square Error(Lower the better)

Multi\_rmserr1 <- caret::RMSE(Multi\_actual\_pred1$predicted,Multi\_actual\_pred1$actual)

Multi\_rmserr2 <- caret::RMSE(Multi\_actual\_pred2$predicted,Multi\_actual\_pred2$actual)

#Mean Absolute Percentage Errors(Lower the better)

Multi\_mape1 <- mean(abs(Multi\_actual\_pred1$predicted - Multi\_actual\_pred1$actual)/Multi\_actual\_pred1$actual)

Multi\_mape1

Multi\_mape2 <- mean(abs(Multi\_actual\_pred2$predicted - Multi\_actual\_pred2$actual)/Multi\_actual\_pred2$actual)

Multi\_mape2

#Cross Validation of Linear Models 1 & 2

Multi\_cvResults1 <- cv.lm(data = Transportation\_df,form.lm = formula(NY.Conventional.Gasoline.Spot.Price....gal.~WTI.Crude.Oil.Spot.Price....barrel. + Brent.Crude.Oil.Spot.Price....barrel.),

m = 3,dots = FALSE,seed = 29,plotit = c("Observed","Redidual"),

main ="Small symbols show cross-validation",legend.pos = "topleft",

printit = TRUE)

attr(Multi\_cvResults1, 'ms')

Multi\_cvResults2 <- cv.lm(data = Transportation\_df,form.lm = formula(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.~Brent.Crude.Oil.Spot.Price....barrel. + WTI.Crude.Oil.Spot.Price....barrel.),

m = 3,dots = FALSE,seed = 29,plotit = c("Observed","Redidual"),

main ="Small symbols show cross-validation",legend.pos = "topleft",

printit = TRUE)

attr(Multi\_cvResults2, 'ms')

#################### RANDOM FOREST REGRESSION ####################

#Applying Random Forest Regression to the training dataset and predicting using testing data

randomFG <- randomForest(NY.Conventional.Gasoline.Spot.Price....gal.~.,data = trans\_train[,-1],importance=TRUE,na.action = na.omit)

round(importance(randomFG),2)

predG <- predict(randomFG,newdata = trans\_test[,-2])

rmseG <- RMSE(predG,trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.)

rmseG

mseG <- MSE(predG,trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.)

mseG

corG <- cor(predG,trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.)

corG

plot(randomFG)

#Random Forest training for Diesel

randomFD <- randomForest(NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.~.,data = trans\_train[,-1],importance=TRUE,na.action = na.omit)

round(importance(randomFD),2)

#Predicting Diesel Price Using Testing Dataset and Finding RMSE, MSE Values

predD <- predict(randomFD,newdata = trans\_test[,-3])

rmseD <- RMSE(predD,trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.)

rmseD

mseD <- MSE(predD,trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.)

mseD

#Correlation Between the Prediction and the Actual Diesel Price

corD <- cor(predD,trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.)

corD

#Plotting the Random Forest object (Error vs Number of Trees)

plot(randomFG)

#Random Forest Cross-Validation for Training and Testing Dataset

resultG1 <- rfcv(trans\_train[,-1],trans\_train$NY.Conventional.Gasoline.Spot.Price....gal.,step = 0.5)

with(resultG1, plot(n.var, error.cv, log="x", type="o", lwd=2))

resultG2 <- rfcv(trans\_test[,-1],trans\_test$NY.Conventional.Gasoline.Spot.Price....gal.,step = 0.5)

with(resultG2, plot(n.var, error.cv, log="x", type="o", lwd=2))

resultD1 <- rfcv(trans\_train[,-1],trans\_train$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.,step = 0.5)

with(resultD1, plot(n.var, error.cv, log="x", type="o", lwd=2))

resultD2 <- rfcv(trans\_test[,-1],trans\_test$NY.Ultra.Low.Sulfur.Diesel.Spot.Price....gal.,step = 0.5)

with(resultD2, plot(n.var, error.cv, log="x", type="o", lwd=2))

######################## TIME SERIES ANALYSIS ########################

Transport.df <- temp

df <-ts(Transport.df$NY.Jet.Fuel.Price....gal.,frequency = 12,start = c(2006,1))

plot.ts(df)

sm <- ma(df, order=12)

lines(sm, col="blue")

k<- SMA(df,n=10)

plot.ts(k)

#Using decompose function to split the time series components into season, trend and irregularities of the seasonal data

decom <- decompose(df)

decompose1 <-df- decom$seasonal

plot(decompose1)

forecast <- HoltWinters(df,beta=FALSE,gamma = FALSE)

forecast$fitted

forecast$SSE

plot(forecast)

future <- forecast:::forecast.HoltWinters(forecast,h=9)

future

forecast:::plot.forecast(future)

plot(future$residuals)