



*Act Locally, Impact Globally*

# FRONTIER TECHNOLOGY INSTITUTE

## DATA SCIENCE CERTIFICATION

MODULE -V: Regression and Time Series Forecasting  
LAB- EXAM (80 Marks, 30 Points)

**You are given a dataset containing crude oil prices. You are required to do the following:**

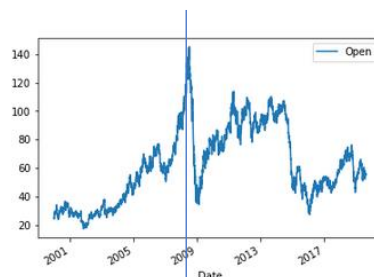
### **Regression: (30)**

1. Upload the dataset in Jupiter Notebook. Use python commands to describe the dataset (info , describe, data types, len(data)). Also check for null values
2. The crude oil prices for Saturday and Sunday are missing. Fill in the missing dates and use the previous day values against Missing days. For example, for Saturday you will add a new date after Friday and use all the values of Friday as the values for Saturday. ( You will have to google it to get the solution)
3. Construct some more features like Year, Month, Week, and Weekday to have some more features.
4. Make the price columns as target.
5. Remove K from **volume** values and set its type as numeric / float etc.
6. Remove Open, High and Low columns from the dataset
7. Use remaining columns as features, price as target. Split into train / test ( 70:30)
8. Apply Multiple Linear Regression , SVR Regression and RF Regression on the data set
9. Compare the Results based on MAE, MSE

**Submit Notebook as well as document containing results summary.**

### **Time Series Forecasting: (50)**

1. Use the above dataset make a series based on price columns by setting date as index ( Missing Days data should be incorporated )
2. The trend line is not uniform throughout the period. Use the data upto 2008 only so that you have a single uniform trend in the data.





FRONTIER TECHNOLOGY  
INSTITUTE

*Act Locally, Impact Globally*

3. Use Mean and variance at different intervals of time series to check whether it is stationarity or not.
4. Use ADF test to check whether Time Series is stationary or not.
5. Is there a trend in the time series?
6. Is there a seasonality in the time series?
7. Try out Power transforms ( Log, Square Root , Inverse ) using BoxCox Transformations . After each transformation check which results in improvement in ADF score. Stay with one which is better according to ADF.
8. Apply moving average smoothing , again check the ADF score
9. Use differencing. Again, check ADF score and stationarity
10. Use 70/30 split and implement and AR Model. What is the RMSE?
11. Implement MA model on 70/30 split. What is the RMSE for persistence model ? What is the RMSE for MA Model.
12. Plot ACF and PACF to choose p and q. Implement ARMA (p,q). What is the RMSE?
13. Implement ARIMA (p,d,q) using proper parameter values for p, d and q.
14. Implement Grid Search ARIMA (p, d, and q)? Mention the grid values you have used for p, d and q. Which is the best combination of parameters? What is the RMSE?

**Submit Note Book Only and provide analysis in the notebook.**

**Deadline: 08 March 2020.**

**Note: Late submissions will be penalised.**