FRONTIER TECHNOLOGY INSTITUTE

Act Locally, Impact Globally

FRONTIER TECHNOLOGY INSTITUTE

DATA SCIENCE CERTIFICATION

CAPSTONE PROJECT

Project Title : Predicting Crude Oil Prices Using Time Series Forecasting and Regression Analysis
Project Manager : Usman Ali
Project Advisor : Dr.Tariq Mahmood
Team Members : Ali Ibad Brohi , Moiz Ali , Yamna Tahir, Khoulah Afzal Qamar

Project Description :

It is a well-known fact that any minor or major fluctuation in the price of crude oil affects the economy at a global level, particularly countries involved in import and export of crude oil. Products related to crude oil (e.g., petroleum) are also important commodities globally. However, crude oil price has a high volatility level, primarily due to the impact of exogenous variables on this price, e.g., stock and trade market fluctuations, demand and supply KPIS, economic indicators, climate, foreign policy and citizen demographics. Most of these variables are out-of-control, and hence, a consequent practice for economists and financial analysts is to gather information about these variables and model it collectively in order to understand the inherent values and patterns.

This activity is complicated: any change in the values of the variables has a significant impact on the crude oil price. This becomes critical for the petroleum use case: price of crude oil contributes more than 50% to the average petroleum price. Hence, it becomes important for analysts to model the irregular and complex effect of the exogenous variables on crude oil price movements.

Traditional Statistical Approaches:

Financial analysts have previously resorted to several statistical approaches to predict crude oil prices. Some important algorithms are as follows: Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, Naive Random Walk, Belief Network (BN) models, Relative Stock Model (RSTK), Naïve Autoregressive (NAIV), Modified Alternative (MALT) model, and the Qual VAR model. Although there are many successful research works, these algorithms are still incapable of comprehensively modelling the complex landscape of crude oil price prediction in real-time.

In the last decade, the advent of Machine Learning (ML) and Deep Learning (DL) technologies has highlighted these limitations of statistical models. ML models can address them by discovering the complex correlations between exogenous predictors and crude oil prices in a more through manner and hence generating predictions that are more accurate.

Modern Learning Approaches:

In fact, we can consider the first algorithm of the "learning" era to be the well-known ARIMA model, which analysis can decompose separately into AR, MA and ARMA models as well. ARIMA, and its seasonal variant SARIMAX, are state-of-the-art approaches to predict any time series data, both in academic research, industrial research and crude oil market application domains. In parallel, we also have Single Exponential Smoothing (SES), Holt's linear trend (HLT), Holt's winter seasonal model (HWSM), along with VAR and VARMAX models for forecasting multiple time series simultaneously. Other famous ML/DL approaches are regression, artificial neural networks, recurrent neural networks, and deep belief networks.

Problem Statement:

Due to the significant impact of crude oil prices on the global economy and petrol consumption, it has become critical to predict or forecast these prices in order to develop appropriate economic plans in advance. We want to use time series forecasting and other related approaches to predict crude oil prices (in PKR/barrel) and then make conclusions regarding the impact of these forecasts on our economy.

Project Requirements:

You are required to execute the following steps and submit all answers compiled in one PDF document. You are also required to submit your Jupyter Lab notebooks.

1. Understand the science of crude oil price movements by watching relevant YouTube videos and reading blogs [some short-duration sample videos: https://www.youtube.com/watch?v=srgUddmiTpU, https://www.youtube.com/watch?v=JJmEQ1mIOPI, https://www.youtube.com/watch?v=rOWlLDVr6eA, https://www.youtube.com/watch?v=wDfJJQHH3Yk, https://www.youtube.com/watch?v=-AjHxUQO2to, https://www.youtube.com/watch?v=-AjHxUQO2to and some sample blogs: https://en.wikipedia.org/wiki/Price_of_oil, https://www.thebalance.com/how-are-oil-prices-determined-3305650, and

https://www.investopedia.com/articles/economics/08/determining-oil-prices.asp and Research Paper .

2. Make your own notes while watching each video or reading each blog. Now, use all your notes to write down a 1-2 page summary of how the crude oil market moves, factors that affect it, and possible implications of these movements particularly on Pakistan [you can do some more research if you want, for instance, go through these links: http://www.pakistaneconomist.com/2018/04/30/rising-oil-prices-impact-pakistan/,https://mpra.ub.uni-uenchen.de/55929/1/MPRA_paper_55929.pdf, https://ijbssnet.com/journals/Vol_2_No_17/29.pdf.
3. You can find the Crude Oil WTI Historical Data at https://www.investing.com/commodities/crude-oil-historical-data . The data is given in US $. There is another dataset at https://www.indexmundi.com/commodities/?commodity=crude-oil&months=240&currency=pkr about Monthly Crude Oil prices in PKR. We need a dataset of daily crude oil prices in PKR. Use the above two links to form this dataset. You can search other sources as well. The dataset should be in from 2000 – Feb 28, 2020.
4. For the provided dataset:
    a. Clean the data if needed
    b. Apply descriptive statistics and interpret your results
    c. Detect need of differencing and if true, difference along with presenting proof
    d. Split into train, hold-out and test sets
    e. In training data:
        i. Seasonally decompose each time series and interpret the results thoroughly
        ii. Determine ACF and PACF plots and hence determine the values of appropriate parameters for ARIMA-based models
        iii. Determine whether we need AR, MA, ARMA, ARIMA, or SARIMAX
        iv. Apply the selected algorithm; notwithstanding information from ACF and PACF plots, we can still try to find out the ideal parametric values either by trying out different values manually, or by using Grid Search in Python

v. Determine the best parameter values and hence, the best forecasts

vi. Interpret thoroughly ALL the model parameters in your own words

vii. Prove this through best performance on hold-out and test sets and show graphs of the fitted model along with relevant performance metrics

viii. Now, forecast for the next five time steps into the future (show forecasts on graph as well as the actual

5. Develop your own strategy to answer the following questions:
    a. Apply Support Vector Regression on your time series and compare results of Steps e(vii) and e(viii) above (with hyper-parameter tuning)
    b. Apply Multi-Layer Perceptron on your time series and compare results of Steps e(vii) and e(viii) above (with hyper-parameter tuning)
    c. **Optional** given the resources required: Apply Deep Learning's LSTM algorithm on your time series and compare results of Steps e(vii) and e(viii) above (with hyper-parameter tuning); this could several hours if GPUs are not available
    d. What can you generalize about the predictive accuracy of different algorithms experimented above?
    e. Based on background knowledge acquired in (1) and (2) above and your overall results, what business-level conclusions can you reach regarding the impact of crude oil price movements on Pakistan's economy in the future.

Cheating or Plagiarism Policy:

Any attempt to plagiarize online content or copy programming code or approach will lead to disqualification from the capstone project.

BREAKDOWN

A high-level breakdown is as under:

1. You are expected to complete Task 1,2 and 3 in first week
2. Task 4 in Week 2&3
3. Task 5 in Week 4

FURTHER DETAILS:

Further details may be shared later on if needed.