Crude Oil Price Prediction using Bayesian Networks

Jeet D. Shah(BTech. ICT), Shantanu Joshi(BS Computer Science)

***Abstract*—This paper focuses on Crude Oil price prediction and regime detection using probabilistic graphical models. Crude Oil plays a crucial role in global economy and hence is a very critical indicator of industrial growth. Crude oil price depends on various macroeconomic, technical and financial factors. To take into account this causality, this paper uses probabilistic graphical models to learn the structure of the crude oil market. This paper proposes condensing data of numerous Crude Oil factors into a graphical model in the attempt of creating a accurate forecast of the price of crude oil and define an accurate trading strategy for the market players. This paper compares the structure learnt by 2 different structure learning algorithms over 3 different scoring methods in order to find the most accurate structure. Secondly, based on the structure learnt, it predicts the behaviour of the oil market.**

***Index Terms*—Crude Oil, Belief Network, Structure Learning, Parameter Learning, HillClimb,K2,Bdeu,BIC,Chow Liu**

I. Introduction

A. *Motivation*

Machine Learning, Artificial Intelligence and Data analytics have played a crucial role in in Quantitative and Computational Finance lately. An increasing number of hedge funds such as Man Group, Two Sigma Investments, Winton Capital, Renaissance Technologies and a number of investment banks such as Goldman Sachs Asset Management(GSAM) and Bank of America Merrill Lynch have been incorporating these technologies in their trades. Hedge funds employing a global macro strategy observe minute changes in the macroeconomic behaviour, in which the price of crude oil is a vital key player. Crude oil plays a key factor in the macroeconomic stability given not only its utilisation in conventional fuels but also its utilisation in creating infrastructure, such as the use of bitumen in laying roads. Therefore, in the long-term, crude oil may heavily influence the rate of economic growth of countries, especially those relying on oil imports and hence have a heavy influence on the performance of the global financial markets. There is widespread agreement that unexpected fluctuations in the real price of crude oil are detrimental to the welfare of both oil-importing and oil-exporting economies. Reliable forecasts of the price of oil are of interest for a wide range of applications [2]. Today, with exponentially increasing amount of datasets, computational power and a plethora of applications of crude oil price forecasts, it acts as enough motivation for us to use Bayesian Analysis for financial forecasting.

B. *Objective*

This paper has two major objectives:

*1) **Structure Learning**:* We have identified various Technical, Economical and Financial factors that affects the crude oil price on a macroeconomic level. We have acquired this time series data from different sources and processed it into a single data frame. We have used this data to define the causality between these factors using correlation and and three different combinations of structure learning algorithms and scoring methods. These are:

- Algorithm: Hill Climb Search; Scoring Method: K2
- Algorithm: Hill Climb Search; Scoring Method: BDEU
- Algorithm: Chow Liu

This paper compares the performances of the above three models.

*2) **Crude oil price prediction**:* From the above section, the model which shows the best performance has been used for predicting the behaviour of the oil market which can be used to decide a trading strategy for the market players.

C. *Related Work*

Research in Ref.[8] used a hybrid model for crude oil price prediction that used complex networks and LSTM model. The complex network analysis tool called the visibility graph is used to map the dataset on a network and K-core centrality was employed to extract the non-linearity features of crude oil and reconstruct the dataset. Then LSTM was used to reconstruct the data. Author in Ref.[1] Used Bayesian Networks to predict WTI crude oil prices. The research paper used many factors such as inflation ,CPI(Consumer Price Index), Supply and deman of OECD and OPEC Countries. These factors were used as nodes in the belief network. The author used a ready made belief network, added some of their own variables and constructed the model using hill climb search structure learning algorithm. Research in Ref.[7] Used Fundamental Factors such as: Value:PE RATIO, PRICE to SALES, PRICE_CASH, Ebitda per Share , Profitability: Ebitda Margin and Buy-Back Yield Sentiment: Volatility and Put/Call Ratio. Authors of Ref.[3] Used belief nets in HFT(High Frequency Trading). Prediction of FX rates is addressed as a binary classification problem and used a Bayes net to solve the problem They have used these to create a Bayes Graph which will predict the movement of SP500 Index. The performance of

This paragraph of the first footnote will contain the date on which you submitted your brief for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Department of Commerce under Grant BS123456.”

The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

these classifiers is compared to that of a dynamic Bayesian network by using real time foreign exchange rates. Research in Ref.[4] conclude that Regression analysis is one of the most common econometric tools employed in the area of investment management. Authors of Ref.[9] this paper adopts a two-stage hybrid model that integrates Deep Learning and Support Vector Machine as a FDP modeling method. Local receptive fields is a technique used in order to select the nodes for each layer of our deep network. Authors of Ref.[10] used DBN(depp belief networks) to forecast exchange rates. Experiments indicate the DBN is better in forecasting than FFNN(feed forward neural network)

II. Dataset

As mentioned earlier, we have taken broadly three categories of data;

A. *Financial Data:*

We have taken factors like the market indices in determining the fluctuation in the price of Crude Oil. We took:

- SP500 Index(GSPC)
- Oil Volatility Index(OVX) Oil Price volatility refers to the degree to which prices rise or fall over a period of time. When relatively stable prices prevail, the market is id to have low volatility.
- Dow Jones U.S. Oil Gas Total Stock Market Index(DWCOGS)

B. *Economic Data:*

Usage of Quandl: Quandl is an API which allows a user to download any financial data which is free to use. Quandl is a platform that provides its users with economic, financial and alternative datasets. We can also sell our data to quandl. We are also taking data from FRED-API. FRED is the Federal Reserve Economic Data and is an open data source service offered by US Federal Government. We have used quandl-API to get data series from FRED. The following factors were taken:

- Consumer Price Index(FRED/CPIENGSL) The CPI Index will weigh the inflation of USA. It measures the average change in prices over time that consumers pay for a basket of goods and services. More the CPI less will be the value of money you have to buy any goods and services as they become expensive(inflation).
- Industrial Capacity:
  - Mining
  - Quarrying
- Oil and Gas Extraction(FRED/CAPG211S)
- Advance Retail Sales: Retail Trade(FRED/RSXFS)

C. *Technical Data:*

We have taken:

- Relative Strength Index RSI(14) What does it tell. The relative strength index will tell us the price strength of an uptrend and downtrend of a security or a stock. If the RSI is increasing then the market is in overbought zone(above a certain threshold) else if it falles below a certain threshold the market is in oversold zone and the stock should be bought for any anticipated uptrend.
- Moving Average Convergence Divergence MACD(21,29,9) It is designed to reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price. MACD is found by subtracting the 26-EMA from the 12-EMA
- 50 Day period Exponential Moving Average EMA(50)
- Average True Range(ATR) indicator ATR Bands show trend in price movements.
- Average Directional Movement Index(ADX) The ADX indicator is an average of expanding price range values. The ADX is a component of the Directional Movement System. This system attempts to measure the strength of price movement in positive and negative direction using the DMI+ and DMI- indicators along with the ADX.

D. *Data Discretization*

We will find the difference in quantities of all the variables on our dataset.

And then set the values=1 if the change is positive else values=0 if change in the next month is negative. Why do we discretize the data? We do that simply because unlike an LSTM/Machine Learning Model Bayes Nets need discrete data to play with. We cannot use continuous data since each node in the graph will have discrete states.

Example: the price difference for Crude Oil Price for first month-(2007-07-31) is 67.5774-61.508=0.0690

1=Bull(upside/uptrend) 0=Bear(downside/downtrend)

Python snippet:

$data_{d}iff = data.diff()[1 :]$

$data_{d}iff1 = np.where(data_{d}iff > 0, 1, 0)$

III. Understanding our Belief Networks

We have used different Structure Learning algorithms to create a belief network best suitable for out project. We have used Hill Climb Search Structure Learning Algorithm to learn the structure of the algorithm and we have also used Chow Liu Algorithm to compare which one is better. Root Node for Chow Liu is the Crude Oil Price.

Moreover We take different scoring methods like K2,BDEU as well and try different combinations with the structure learning algorithm to get the best fir for our model.

A. *Structure Learning*

- **Hill Climb Search:** Score+search method: In this algorithms a function f_s used to score a network/DAG (Directed Acyclic Graph) with respect to the training data,and a search method is used to look for the network with the best score. Unlike other algorithms, instead of restricting the search space, hill climb search takes advantage of the computations carried out at each search step to guess which edges should not be considered from then on. In this way, the search space is pruned progressively as the search advances. [6]

- **Chow Liu:** The goal is to find a tree that maximizes the likelihood of the training data.
  Algorithm:
  - compute weight I(X_i, X_j) of each possible edge (X_i, X_j)
  - find maximum weight spanning tree (MST)
  - assign edge directions in MST

B. *Parameter Learning*

Parameter learning is of two main types:

- **Maximum Likelihood Estimation:** Maximum Likelihood Estimation is a framework for solving the problem of density estimation. It involves maximizing a likelihood function. It provides a framework for predictive modeling in machine learning where finding model parameters can be framed as an optimization problem.
- **Bayesian Estimation:** The Bayesian Parameter estimator begins with already existing prior conditional probability tables that express our beliefs about the variables before the data was observed.

IV. Results

A. *Structures Learned*

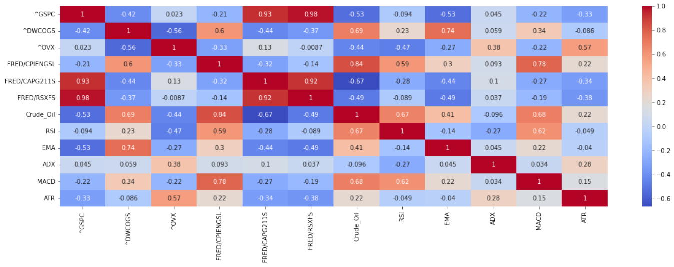

Fig. 1. Our Correlation matrix to understand which variables are related with each other to get a rough idea of what we can expect. We can verify the edges of the structure learnt from the correlation between them.

Learning Algorithm	Scoring method	Accuracy
Hill Climb Search	K2	**70.8333%**
Hill Climb Search	BDEU	**70.8333%**
Chow Liu	-	54.1667%

TABLE I
Table for comparison

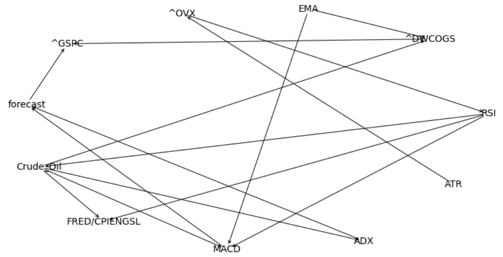

Fig. 2. Our Bayesian Network for the financial market using Hill Climb Search,score=K2

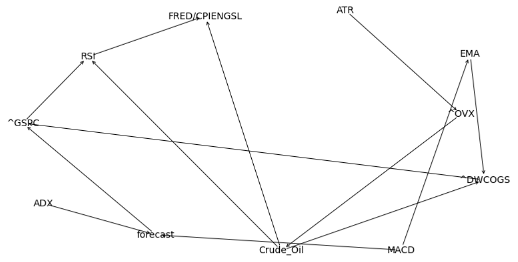

Fig. 3. Our Bayesian Network for the financial market using Hill Climb Search,score=BDEU

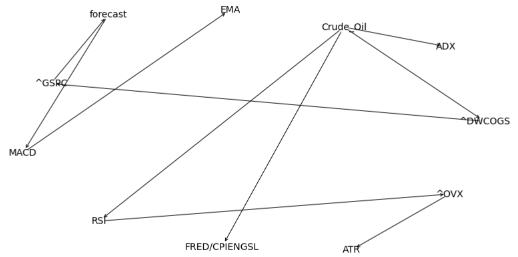

Fig. 4. Our Belief Network using the Chow Liu Algorithm, Root node = Crude Oil Price

B. *Regime Detection*

We have used our model to detect the regime also, Like when the regime is bull then it is an uptrend in the oil prices which indicates a Long Position whereas a bear would indicate a Downtrend and a short seller would find this opportunity to sell using futures and options. Moreover a sideways market(stagnant) market is a market where neither bull and bears have the upper hand but this period is generally a period where there is buying and selling happening at the same rate. This graph is a great way for any trader/investor to get the performance of any security over the years and is easily understood using color of regimes.

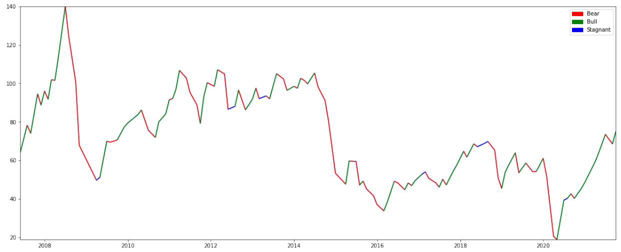

Fig. 5. Our Final Plot of the Model, which predicts the regime(bull,bear and stagnant) of the Oil market over the years.

V. Conclusion and Future Work

This paper contributes to the existing literature in a number of ways. Very less work has been done in studying the causality between various macroeconomic factors affecting the oil market. [1] has done work in the above direction but, it does not take into account the diverse factors as this paper does. It also uses only one learning algorithm for structure learning and does not compare performances of other algorithms. Moreover, above all, we have been able to achieve an accuracy of nearly 71% as compared to 41% in [1]. 71% in itself is a great performance from machine learning perspective in general. Based on our predictions, investors can safely invest in the market and make profit with 71% certainty.

We can improve this performance by using LSTM to model the time series data better and better estimate the missing values in the data. Deep learning models can be used for better parameter estimation. That can significantly improve the results.

References

[1] , D. A. (2018). (PDF) application of probabilistic graphical models in ... Research Gate. Retrieved October 22, 2021, from https://www.researchgate.net/publication/324859915Application of Probabilistic Graphical Models in Forecasting Crude Oil Price.

[2] G. O. Young, “Synthetic structure of industrial plastics,” in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.

[3] S., amp; Stella, F. (2014). A continuous time bayesian network classifier for Intraday FX prediction. Quantitative Finance, 14(12), 2079–2092

[4] Bayesian estimation of stochastic volatility models. (2015). Bayesian Methods in Finance, 229–246. https://doi.org/10.1002/9781119202141.ch12

[5] Bayesian linear regression model. (2015). Bayesian Methods in Finance, 43–60. https://doi.org/10.1002/9781119202141.ch4

[6] Gámez, José and Mateo, Juan and Puerta, Jose, 2011, Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhoo, Data Mining and Knowledge Discovery

[7] , M. (2018, January). Bayesian networks for Financial Market Signals Detection. Research Gate. Retrieved October 12, 2021, from https://www.researchgate.net/publication/323333070BayesianNetworks for Financial Market Signals Detection.

[8] , A. A. (2019, November 27). CPPCNDL: Crude oil price prediction using complex network and deep learning algorithms. Petroleum. Retrieved November 22, 2021, from https://www.sciencedirect.com/science/article/pii/S2405656119301117.

[9] IEEE report (October 20-21)https://ieeexplore.ieee.org/document/7358416

[10] (2011). Forecasting exchange rate with deep belief networks. 1259-1266. 10.1109/IJCNN.2011.6033368.