A Macroeconomic Model for Forecasting Crude Oil Prices with Feedforward Neural Network Grid Search Experimentation

by

Md. Ragib Aunjum
16301096
Muhammad Naqi
16301083
Sifat Jamil
19341029

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
December, 2019

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Ragib Aunjum
16301096

Muhammad Naqi
16301083

Sifat Jamil
19341029

Approval

The thesis titled "A Macroeconomic Model for Forecasting Crude Oil Prices with Feedforward Neural Network Grid Search Experimentation" submitted by

1. Md. Ragib Aunjum (16301096)
2. Muhammd Naqi (16301083)
3. Sifat Jamil (19341029)

Of Fall, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on December 26, 2019.

Examining Committee:

Supervisor:

Mahbubul Alam Majumdar, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:

Mahbubul Alam Majumdar, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Crude oil is one of the most important determinant of the global and national economy and important decision making factors of industrial activities. For this reason, numerous mathematical and machine learning approaches have been conducted to predict the future trend of oil market. Yet, to predict the price of oil is one of the most challenging issues out there because the high volatile nature of oil market and the dependency of price on other factors. In many approaches on predicting oil price use machine learning algorithms, the only factors considered are the opening and closing prices. Thus, the implementations did not reflect the price pattern truly and also hampered the sudden ups and downs of price because the oil market does not only depend on the daily pricing behavior. By reviewing the historical data of oil market it can clearly be seen that the oil market is heavily affected by the geopolitical, technical and macroeconomic factors. For example, geopolitical factor such as war in middle east made the oil price soared high and broke the pattern of daily fluctuations by a large margin. And also, it can easily be seen that the everyday demand of oil along with the quantity supplied affects the oil market. So, these factors along with other macroeconomic and technical issues must be addressed to successfully determine the oil price trend. To justify our claim, we approach to predict the oil price using only the opening and closing market price by ARIMA, SVR and Linear regression model. Afterwards, the macroeconomic, technical, geopolitical factors were considered to predict oil price using Feed Forward Neural Network and compared the results with the ones we have found on the previous models.

Keywords: Crude oil price prediction; Machine Learning; Technical Factors; Macroeconomic Factors; Geopolitical Factors; ARIMA; Artificial Neural Network; Grid search

Dedication

To our loving parents who support us on our ups and downs and love us unconditionally....

Acknowledgement

We are highly grateful to our supervisor Professor Mahbubul Alam Majumdar, Chairperson of the Computer Science and Engineering department, BRAC University for his guidance to complete this thesis. He always managed to give us time on his busy schedule whenever we needed his knowledge, advice and suggestion on both coding part and thesis write up. Without his cordial help, we could not finish our work in this way.

We like to thanks all the respected teachers for the knowledge they provide us throughout our academic journey and our friends and family who aided us to overcome the impediments of life.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN Artificial Neural Network

$ARIMA$ Autoregressive Integrated Moving Average

FNN Feedforward Neural Network

$GARCH$ Generalized Autoregressive Conditional Heteroskedasticity

$LSTM$ Long Short Term Memory

MSE Mean Squared Error

$RMSE$ Root Mean Squared Error

$S\&P500$ Standard & Poor's 500

SVR Support Vector Regression

Chapter 1

Introduction

In today's world, crude oil is a prime commodity which controls the geopolitics and economy. A simple change in the oil price has far reaching impact on the global economy and politics and there is no denying that almost all the countries are heavily reliant on the oil for its economic development and energy efficiency. The oil dependency of world economy was predicted in the mid-80s when the Arab world was in a serious turmoil, the oil price rose up significantly and the world realized how important the security of oil means. Since then the oil price is at the center of the attention of the global politics and the leaders around the world is eyeing the oil rich countries to ensure that they can stay on the smooth economic bandwagon. According to Balke and Yücel, (2002) the rising oil price causes significant damage to the US economy than the falling oil price to stimulate it [12]. According to Lee et al. (2001) Oil price has significant impact on the monetary policy of Japan [11]. All the developed country is facing similar situation in the fluctuation of oil price, hence the demand for oil is always at the rise. In retrospect, overall demand for oil in terms of volume has decreased, but the number of user of oil has increased. Every passing year the countries around the world is trying to find an alternative to oil, but the overall demand for oil is still at rise as all the countries face higher ownership of private vehicles and rapid growth of industries around the world is at the high at this moment. The reserve of the crude oil in the surface is finite and there is still no estimation of how much oil is undiscovered beneath the earth. This uncertainty put oil at the center of attention in the world and the oil price becomes the center of all economic activities.

The change in oil price has significant impact in our daily lives in a number of different ways, if we want to see the effects in macroeconomic view we can understand how vulnerable we are in the fluctuations of the price of the oil. Although it is really difficult to find out precise impact of oil price shock in macroeconomics, there is no denying significant impact (Bernanke and Watson, 2004) [13]. Undoubtedly, the production of energy is directly connected to the production of crude oil. Thus, the fluctuations in oil market impact the lives of the population on a large margin. Moreover, in many industries oil is used as a raw material, so the price of raw materials goes up with the increase of the oil price, resulting in higher production cost. Higher production cost leads to price hike of daily commodities and the impact is all on the general consumers. The poor and middle class citizens become the victim of the rise as their means for survival gets affected if the price of commodity rises along with the energy price. When the effect is directly on the macro economy the micro economy will be effected as well. Macro indicators of the economics such as employment rate, inflation, trade balance, exchange rates and stock market price usually get negatively affected when the price of oil rise up. When the negative flow in the market is prevalent, the spending power of people fall down and demand for products goes down subsequently. On the other hand, when the oil price plummeted the reverse situation occurs.

The effect of oil price has significant impact on the exchange rate of Dollar. It is found by Lizardo and Mollick (2010) that, when the price of oil is added to the monetary model of the exchange rates, the change of the US dollar price is elucidated by the change of oil price [25]. In a Dollar based global economy the exchange rate fluctuation brings major changes in any countries economy, the poor countries who are not considered a significant player in the world politics suffers as there is no way to do anything about the situation, in many cases these countries need to borrow a significant amount of money from various international financial agencies to make sure the import related daily commodities that are necessary for their citizens are available. Because of this exchange rate constraint, the export is badly affected as well as there are some countries who are exclusively dependent on export of only one product to fuel the economy. Countries like Bangladesh where the main export is ready made garment the exchange rate is of crucial importance as the export order and the supply of the good need to be calculated on the rate of exchange. The rate of exchange can cripple a country's economy, countries who are under the sanction of United States of America is finding it difficult to survive. Countries like Iran is facing major problems regarding the rate of exchange of USD because its fluctuation is reflected on the inflation in the country. According to Farzanegan and Markwardt (2009) the Iranian economy is facing serious difficulties to maintain the daily needs of the citizens due to the sanctions and the exchange rates [20]. There are so many other nations where this situation is present because of the exchange rate issue.

The price of oil made dramatic impact on the agriculture and agricultural goods. According to Wang and Yang (2014) around 20 to 40 percent of the variations in the agricultural commodity price depends on the oil price [33]. Most of the farmers around the world use oil for the irrigation purpose. As a result, the production cost for crops grow with the oil price rise so the farmers need to sell the products in the market in a higher price. Ji and Fan (2012) found that there are some crops which are directly affected by the oil price alteration [27]. Crops like cotton, soybeans, corn and the crops in the same category is effected by the change in the price of oil. The agriculture sector uses heavy machineries these days and these machineries are using some form of crude oil, the tractors and the cutting machines and all the engine based vehicles used in the agriculture is using the oil on a regular basis. Operating these vehicles and equipment are now a necessity in agriculture sector and the price of oil will have adverse effect on the overall agricultural sector as a whole. Liu (2014) finds that energy price constitutes to almost 15 to 20 percent of the agrarian cost in US [30]. The economic implication in the agricultural sector in respect to oil price can be illustrated by the high production cost along with high product price resulting in increase in agricultural products price. With the upward trend of price, the demand for some product will go down and the surplus will be available, on the other hand the rising price of the commodities will make farmers to cut down the production. If the production decreases the number of labor will be less and there will be some unemployment in the agriculture sector. On the other hand, if some oil rich country can produce the crops in a lower rate, the government may try to import the crops for abroad creating tension in the agricultural industry. There are far reaching effect of the price of the oil in agricultural sector apart from that aspect as the rise in oil price may destroy some market of agricultural goods in some country. At the same time oil price effects, the food industry as well and making the life of the people at risk if the price rises. But the fall of price will never create same effect as recovering from the adverse situation is always difficult.

Price of the crude oil has impact on the food production sector as well, as it is observed that there is a rise in the price of the food items when the price of the oil surges up. According to Diao et al. (2008) most of the projections relating to the food price concluded that the food prices will stay relatively high for many years to come as the international demand for the food product is rising, one of the reasons for the rise in international food demand is related to the increasing population in the world [19]. The food market around the world is really volatile in midst of changing climate and the scarce resources to grow foods. The food industry is exclusively depended on the oil on other aspects as well as we can see the use of oil in almost all means of transportation in the industry. The economic implication of the rise in food industry can be shown in the fall of demand of certain food items, if the price rises the demand falls and the company and the population face difficulties if the oil price is rising. On the other hand, the decrease in oil price has immediate impact on the food sector as well, as the fall in price of oil will make the price of the food lower, he demand will go up and in order to cope up with the demand more and more companies can come in the food industry. In the process there will be employment opportunity, more options for the consumers and the economy will go up as the new resources has been developed. The far reaching effect will be beneficial for the country and the economy of the country in general.

As it is indicated that the unemployment rate is related to oil price, it can be considered as a prime determinant of the economic development of a country. According to Van Wijnbergen (1985), there is significant relationship between unemployment and oil price [2]. Rafiq et al (2009) states that, there is strong evidence on the impact of the oil price volatility and macroeconomic factors, such as unemployment and investment over the period from 1993 to 2006 [23]. The unemployment rate indicates the health of any economy and the political stability of the country. The stability of a country is greater if the unemployment rate is minimal, but the upsurge in the oil price can create tensions in so many different sectors where it becomes difficult to sustain better employment in all the sectors. Shaari and Abdul Rahim (2013) find that there are positive and negative aspects in Malaysian economy with the fluctuations of oil price. If the oil price is going towards positive direction all the industries find it difficult to produce as much product as they can do before as a result jobs are cut to maintain low operating cost [29]. If we look at Bangladesh for example we can see the similar trend, our ready-made garment industry is heavily dependent on the oil for the production as the power outage is common in our country if the price of the oil is on the rise the production cost becomes higher and the overhead cost goes significantly up, in this kind of situation the company have no option but to lay off some employees to maintain the balance and produce according to the cost. The unemployment rate is one of those macro-economic factors which shows the overall situation of a country, with direct relation to the income of the people, unemployment plays great role on the domestic product consumption and import of product. As there are so many factors related to the unemployment and a direct connection amongst the oil price fluctuation and unemployment problem is found, it is of pivotal importance to speculate the price of the oil and have some precautionary plans from the government to make the situation better for the country and save the population from the adverse effect of the fluctuation of oil price.

To save the population from oil price fluctuation the governments usually carry a subsidy program to ensure stability of the economy. Dick (1980) finds that oil subsidy is a constant feature in the Indonesian economy [1] and Jbir and Zouari-Ghorbel (2009) found similar policies in Tunisian economy as well [22]. Sterner (1989) explains there is an implicit subsidy in the oil rich countries of Latin America [15]. As we can see that from Asia to Africa to America the subsidy on the oil is prevalent and the effect of rising price of oil is of significant national interest. Because of the subsidies the effect of small fluctuations in oil price can be mitigated and are never felt by the domestic population, but drastic fall or rise of the price has far reaching effects on the economy of almost all the countries in the world. All the governments of the developed countries in the world has specialized wings who keep a close eye on the oil price through the prediction of oil market which shapes the effectiveness of an economic system in a country.

The history of global politics related to the oil price is not a recent phenomenon and there are ample examples of using the oil market as a weapon against many nations. Although The United nations pleaded that the oil embargo and oil price and production decisions are not their policy (Rothschild, 1975) [16]. Countries are using the oil price to control the global economy, as there is no suitable alternative is available to replace it. Nordhaus (2009) proves that US has been using the embargo on the oil production of Libya and Iran to control these countries [17]. According to Bauer et al (2015) Russia is facing a self-imposed sanction on oil because of various global pressure [18]. Crude oil can be considered the new 'apple' in the discord among the global powers in the present times. Oil price is one of the core drivers that moves the tide of the global politics, economics, production, agriculture and industries. Thus, the importance of predicting the price of the oil is at the top chart of all the governments around the world and it seems to stay so in the recent future.

Chapter 2

Literature Review

The complex nature, dependency on various factors and high volatility makes the prediction of oil price an epitome of choice for the predictors to do. On that path, many steps and procedures have been taken in which some of the knowledge of machine learning have also been implemented. Various algorithmic approaches such as Support vector machine, Recurrent Neural Network, Evolutionary Neural Network, Stream Learning, LSTM and more algorithmic approaches had been done in prediction process. And the factors that considered on most cases are the market demand and supply [4] and the volatile nature of the market [9].

Among the prediction algorithms the Support Vector machine had been profoundly used in different implementation. On their approach Xie et al. (2006), used the WTI crude monthly price from January 1970 to December 2003 with a total observation of 408 [17]. The prediction of the oil price was compared with RMSE and direction statistics to asses their results efficiency. For full period the obtained result of RMSE was 2.1921 and direction statistics was 70.85. On their implementation with SVM, Guo et al. (2012) adopted a model with Genetic Algorithm optimization [21]. Their observation for the traditional SVM implementations found that kernel function parameter and penalty factors are solely dependent on the experience. With the Brent oil price data from December 2001 to August 2011 they tested and trained their dataset against the traditional SVM but not with other algorithms. With a hybrid model of Autoregressive Moving Average and Support Vector Machine HE et al. (2009) predicted the daily oil price from the data dated 2002 to 2008 [21]. Gabralla et al. (2013) approached with ensemble machine learning with Support vector regression, K star and Instance based learning for their model [28].

There were also quite a large number of implementations of Neural Networking in terms of the prediction of oil market. For example, Mahdiani and Khamehchi (2016) used modified Neural Network model to predict the oil price and their main way was to find it through the large amount of monthly and daily price data gathered from WIT database [34]. With a total 219 points of daily oil prices and a total of 95 points of monthly oil prices they made their model with the composition of Artificial Neural Network and Genetic algorithm. Before that implementation, Kaboudin(2001) used GA and ANN for crude oil price prediction with different data points [10]. Also, Shambora and Rossiter (2007) [18] used ANN to predict oil price whilst Mirmirani and Li (2004) used VAR model to predict and used that against the result of ANN and GA combined model [14].

Moshiri and Foroutan (2006) used GARCH and ARIMA model to forecast oil price for the period 1983 – 2003 [16]. Then, for forecasting the series they set up a nonlinear Artificial Neural Network to make better forecast because the chaotic process of oil pricing. They eventually came to discover that ANN worked significantly better than GARCH and ARIMA. Output function for their proposal was –

$$y = F\left[\beta_0 + \Sigma_{j=1}^{q} G\left(xy_j\right)\beta_j\right] \equiv f(x,\theta) \tag{2.1}$$

Abdullah and Zeng (2010) proposed a model using the knowledge of Machine learning with the help of Hierarchical Conceptual model and Artificial Neural Network Quantitative model [24].

On the contrary to the popularly used models, Huntington (1994) implemented a complex electrometric model [5]. Gulen (1998) applied a co-integration analysis [7] e. Lanza et al. (2005) used Error correction model [15] to predict crude oil price.

In lieu of the wide use of the popular algorithm as ANN, SVM, GP have some drawbacks to deal with. For example, ANN sometimes stuck in local minima and experience overfitting. Also, ANN along with SVM, GA are highly sensitive to parameter selection.

According to Rathi and Agarwal (2014) on their research on mammography mass dataset they came to discover that Feed Forward Neural Network performed significantly well than the Artificial Neural Network [31]. The result showed that the FNN scored 90% on accuracy whereas the ANN had 87.5%. At his research Svozil et al. (1997) asserted that FNN is robust and has a graceful degradation over the increasing amount of noise [6]. Because of the scarcity of the approach of Feed Forward Neural Network on oil price prediction along with the performance satisfaction on different occasions we chose to implement our model with FNN. We used various factors including geopolitical, macroeconomic and technical factors correlated with price to implement the model in FNN.

We also used Autoregressive integrated moving average model with closing price data because it has the ability to predict well with large amount of data and it can put time restriction over any variable that has stopped giving values after certain time period [8]. Results from SVR, Linear Regression and exponential model have also compared with the results from ARIMA and FNN output.

Chapter 3

Machine Learning Algorithms

3.1 Linear Regression

Linear Regression shows the relationship between the predictors and the target output. Linear Regression consist of the best fitting line according to the points that can be found by plotting the points [10]. A linear model is a sum of weighted variables that predicts a target output value given an input data instance. Mathematical representation of Linear Regression may include –

Input feature vector: x=(x$_0, x_1, x_2, \ldots, x_n$)
Predicted output: o =w$_0 x_0 + w_1 x_1 + +w_n x_n + b$
Parameters to estimate: $\{w = (w_0, w_1, \ldots, w_n)$ [feature weights/model coefficients] b [constant bias
intercept]

So, according to the equations we can see that we need to find out the weight (w) and bias (b) to reduce the MSE but there is no term to reduce model complexity [11]. The parameters are estimated from data and different methods.

To estimate the success rate of the model we use it to compare with R- squared model which is used as baseline model. R squared take value between 0 and 1 where closer to 0 represents the failure of the model and a closer to one value represents perfectness of the model [12].

R squared can be represented by –

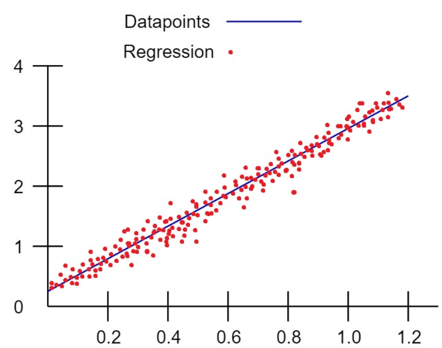


Figure 3.1: Linear Regression

$$R^2 = 1 - \frac{SSE}{SST} \tag{3.1}$$

In this equation, the sum of squared error of regression model and R-squared model are represented by SSE and SST respectively.

$$SST = \sum_{i=1}^{n} (o_i - \overline{o}_i)^2 \tag{3.2}$$

$$SSE = \sum_{i=1}^{n} (o_i - \hat{o}_i)^2 \tag{3.3}$$

3.2 Support Vector Regression

SVR which is the acronym for Support Vector Machine has the following characteristics –

1. Number of support vectors

2. Absence of local minima

3. Usage of kernels

4. Sparseness of the solution

Support vector machines have versatile usage in classification methods and also can be used as regression model. For the regression purpose Support Vector Regression (SVR) is used which has a similar mechanism of support vector machine with some few changes.

To begin with, if we consider the output of SVM than we can see that we get discrete values as we group or classify data into different classes [13]. But on the other hand, SVR predicts continues value as it is used for regression. In that reason, it becomes cumbersome to predict the real value of regression analysis with SVM. To solve the issue, a new term has been used named epsilon that determines the tolerance margin.

In both SVM and SVR the solution is presented by groups of data or subset of training points [14]. And using the epsilon ensures the optimization of reliable generalization bound and presence of global minima.

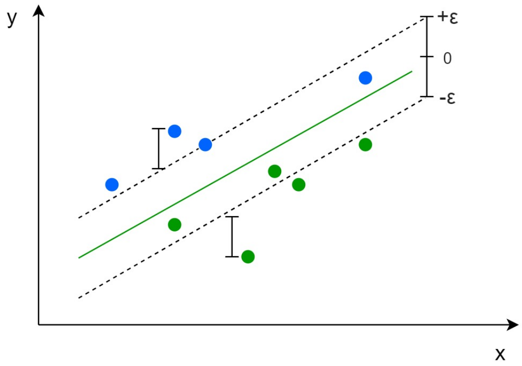


Figure 3.2: Support Vector Machine

All the inputs in SVR are mapped into m dimensional space. Mathematical representation of the feature matrix can be done as follows-

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{m} \mathbf{w}_j g_j(\mathbf{x}) + b \tag{3.4}$$

Here, b = Bias Term
$g_j(\mathbf{x}), j = 1, \ldots, m_{\text{=nonlinear transformation}}$

The performance was measured by Epsilon insensitive loss function which can be represented as -

$$if |y - f(\mathbf{x}, \mathbf{w})| \leq \varepsilon, L_\varepsilon(y, f(\mathbf{x}, \mathbf{w})) = 0 \tag{3.5}$$

$$\text{else }, L_\varepsilon(y, f(\mathbf{x}, \mathbf{w})) = |y - f(\mathbf{x}, \mathbf{w})| - \varepsilon \tag{3.6}$$

SVR can be of two types -

i) Linear SVR
It can be represented by simple mathematical equation –

$$y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b \tag{3.7}$$

ii) Nonlinear SVR
For data separation linearly kernel functions are being used. Kernel functions moves the data to higher space so that the separation is made possible –

$$y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b \tag{3.8}$$

$$y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot k(x_i, x) + b \tag{3.9}$$

Kernel functions:

$$k\left(\mathrm{x}_i, \mathrm{x}_j\right) = \left(\mathrm{x}_i, \mathrm{x}_j\right)^d \tag{3.10}$$

3.3 Autoregressive Integrated Moving Average

ARIMA, which is the acronym for autoregressive moving average is a widely used statistical model to forecast time series data [15]. It can be easier to understand the inner meaning of the model if we just divide the acronym-

AR : AR stands for Auto regression. The name suggests that this model finds out the connection between observations of different time interval.

I : Here I is for integrated. The model makes the time series stationary by differencing raw observations.

MA : MA stands for moving average. It shows the errors from regression which should be considered while predicting time series using this model.

ARIMA model can be of two types-

i) Non-Seasonal ARIMA model: Non-seasnoal ARIMA model is used when the time series does not show any seasonal affects. ARIMA(p,d,q) is the standardized representation of non-seasonal ARIMA model. The parameters stand for –

p : Observation of lag in the model. It helps to adjust the line to be adjusted for time series prediction and also shows the autoregression part of the ARIMA model. Linear regression can be seen in a perfect AR model.

d : The degree of differencing. To work with time series, one needs to make the time series from non-stationary to stationary. The term 'd' represent the order of differencing that is needed for stationary time series conversion. If the time series already shows stationary characteristics than the value for d is 0. The value of d becomes 1 when after first differencing the statistical values of the time series becomes constant. Otherwise, higher order differencing is needed for time series conversion from non-stationary to stationary.

q : It represents the moving average which shows the numerical value of lagged error term and represents the moving average part of the model. These components are not a part of trend or seasonality and must be added or removed from the desired output.

If the two terms become zero, the model can be described as the nonzero acronym of the model such as –

AR(1),I(1) and MA(1) which stands for ARIMA(1,0,0), ARIMA(0,1,0) and ARIMA(0,0,1) respectively.

ii) Seasonal ARIMA model: If the time series has seasonal effects on it where a pattern of ups and downs can be seen on different season of the year, then the Seasonal ARIMA or SARIMA is being used. The Standard representation is ARIMA(p,d,q,m) where -

m : Number of periods

(p,d,q) : Same as the parameters of non-seasonal ARIMA.

In order to find out the value of p partial autocorrelation is being performed and for q autocorrelation is used.

1. Autocorrelation Function : It shows how correlation amongst two variables changes as their separation changes. At lag k it can be expressed by,
k = k/0 where k = cov(yi, yi+k) for any i.

2. Partial Autocorrelation Function : It is the correlation amongst two variables taking assumption that we take into account the values of some other set of variables.

Mathematical denotion –

$$\mathrm{y} \sim \mathrm{t} = \phi 21\mathrm{y} \sim \mathrm{t} - 1 + \phi 22\mathrm{y} \sim \mathrm{t} - 2 + \mathrm{et} \tag{3.11}$$

3.4 Long Short Term Memory

The main disadvantage of RNN was Vanishing and Exploding gradients. To solve this issue, many variations was developed. Most notable of these variations is Long Short Term Memory or LSTM is short. An LSTM network consists of many units and each of these units is responsible for remembering the previous data that have flowed so far and to forget irrelevant information. There are activation layer functions which are called 'gates' in LSTM. Gates do the work of memorizing and forgetting. Each unit also consists of a vector named the Internal Cell State that records the data retained by the previous unit.

There are four different gates:

Forget gate: The unnecessary data in the cell state needs to be removed and the forget door is responsible for this. This gate receives two inputs x_t (input at a certain time) and $h_t - 1$ (output of the previous cell) and then this gate multiplies these inputs with weight matrices and afterwards adds the bias. Then an activation function processes this result and produces a binary result. If the output of a cell state is 0, the stored information is erased and if the output is 1, the stored information is kept for future use.

Input gate: Input gate controls the input of data to a certain cell state. For governing the data, first, the information is processed by a sigmoid function. Then, by using inputs x_t and $h_t - 1$, the values that requires to be remembered are filtered.

Using tanh function, a vector is generated which includes all possible input values. Multiplying the vector values and governed vales produces the desired output.

Input Modulation gate: This is basically a sub-part of the input gate. This is the reason it's not even stated by many literature on LSTM and presumed within the input gate. This gate is responsible for modulating the information that is going to be written in the internal cell state by the input gate. The information is converted to zero mean for faster convergence and better learning time.

Output gate: Output gate is responsible for extracting necessary and relevant information form the cell state gate and presenting it as an output. Tanh function generates a vector, the sigmoid function regulates the information and necessary values are filtered and remembered using input gates. The vector values and regulated values are multiplied to generate the output.

AN LSTM Network's simple workflow is identical to a RNN's workflow. The only difference is that the Inner Cell State is also transmitted with the Secret State.

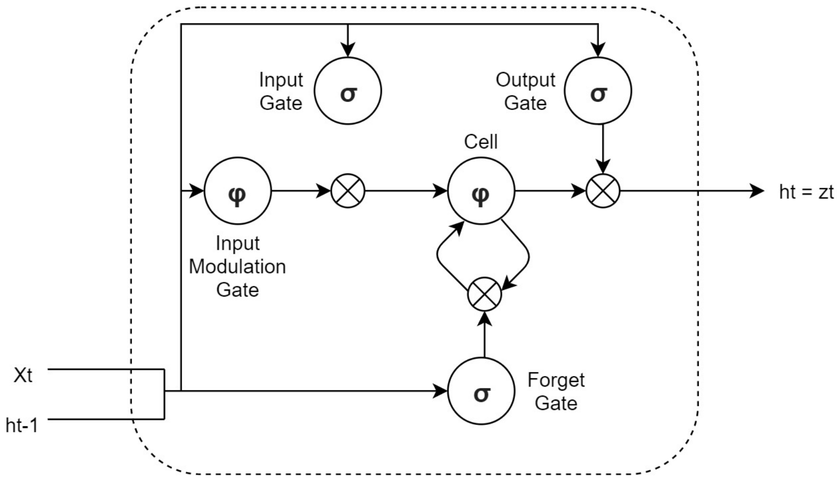


Figure 3.3: Long Short Term Memory

A Long Short Term Memory Network's simple workflow is identical to a Recurrent Neural Network's workflow, with the only exception being that the Inner Cell State, along with the Secret State, is also transmitted.

3.5 Feedforward Neural Network

Deep FeedForward networks are considered as the pioneer of deep learning models. Popular models like RNNs and CNNs are just some modified versions of FFNN..

These networks are called feedforward because of the information flowing in a forward manner (no loops or cycles). For example, a is used to determine an intermediate function in a hidden layer. The function the generates output to calculate y. Here, if the generated output from the last hidden layer can be passed and used by the first hidden layer, the whole thing can be considered a recurrent neural network. Using feedback from a previous operation to achieve more accuracy is the trick here.

Various functions altogether compose these networks. An acyclic graph describes how the multiple functions work together in a group. There can be necessary number hidden layers between the input layer and output layer. These layers can contain any number of hidden units. Each unit works like a neuron because it takes input from the generated outputs of units of previous layers and runs the computation function to get its own activation value. The sequence for the flow of information is input layer to hidden layer to output layer.

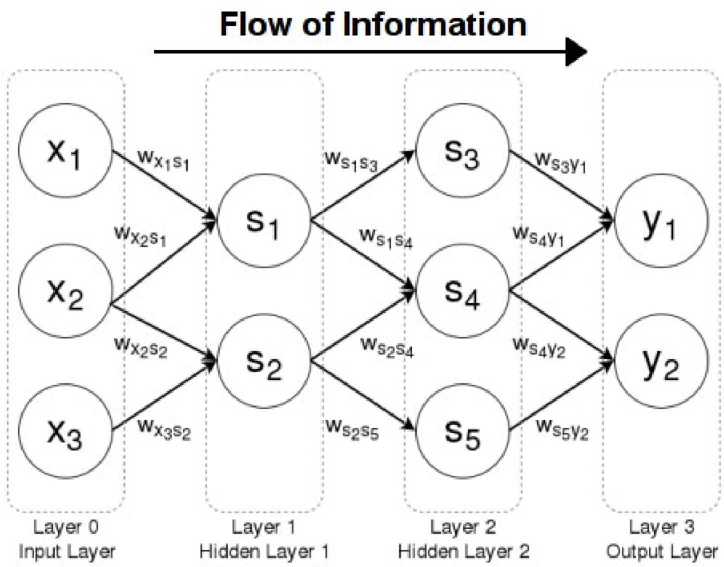


Figure 3.4: Forward flow of information

Neural networks are not limited to only linear functions like linear machine learning models. When the data cannot be represented linearly, linear machine learning models might not perform well or face difficulties to find the approximation while it's an easy task for neural networks.

For designing any feed-forward neural network, we need to choose the ingredients so that the requirements are fulfilled for a better prediction.

Activation Function

Mathematical equations that are used to determine the output of a neural network are known as activation function. Every single unit in the network is attached with this function. This function determines when to activate judging the input's relevancy with the model's prediction.

Activation function needs to be computationally efficient because they are calculated for every single unit which can be millions in number. Need for such efficiency has produced a variety of functions. There are 7 kind of activation function.

For our model, we have used ReLU (Rectified Linear Unit)

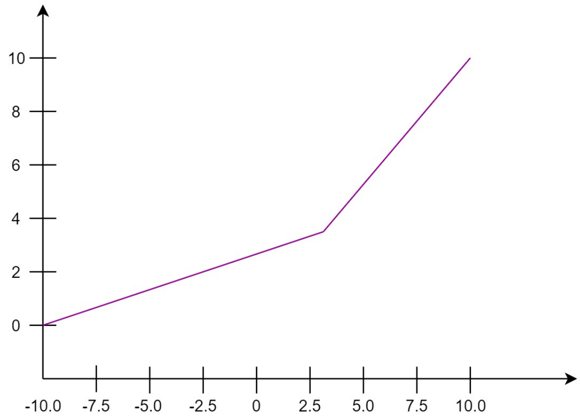


Figure 3.5: ReLU (Rectified Linear Unit)

Advantages

i) Allows the network to converge quickly.
ii) Allows backpropagation and has a derivative function.

Disadvantages

i) If input is zero or less than that, the gradient function disappears and so the network cannot train itself due to the lack of backpropagation.

Optimizer

Optimizer basically indicates the optimization functions responsible for minimizing the cost function. The values of the biases and weights after every cycle of training is updated by these functions and thus global optimum can be reached. There are two types of optimization algorithm:

First Order Optimization Algorithm

Algorithms responsible for minimizing or optimizing a cost function with respect to the parameters by using its gradient values. The First Order derivative tells us if the function at a particular point is decreasing or increasing, in short, it gives the tangent line to the surface.

Second Order Optimization Algorithm

Algorithms that use second order derivatives to minimize the cost function. Second order derivative is expensive to calculate, so not much is used in the second order. We can know if the first derivative increases or decreases the curvature of the function from the second order derivative. Second Order Derivative gives us a quadratic surface that affects the Error Surface curvature.

There are many optimization algorithms like Stochastic Gradient Descent, Adam, Adagrad, RMSprop etc. We have used RMSprop for our model.

RMSprop optimizer restricts vertical oscillation (along the y-axis) and concentrate on which direction is more suitable in the x-axis. This increases the learning rate and enables our algorithm to increase the length of the steps horizontally resulting in a quicker convergence. The following equations show how we calculate the gradients for the RMSprop. The value of momentum is represented by beta and usually set to 0.9. If the value of v_dw comes really close to 0 then the weights might explode. To prevent this, a parameter epsilon is included in the denominator which is set to a small value.

$$\begin{aligned} v_{dw} &= \beta \cdot v_{dw} + (1 - \beta) \cdot dw^2 \\ v_{db} &= \beta \cdot v_{dw} + (1 - \beta) \cdot db^2 \\ W &= W - \alpha \cdot \frac{dw}{\sqrt{v_{dw}} + \epsilon} \\ b &= b - \alpha \cdot \frac{db}{\sqrt{v_{db}} + \epsilon} \end{aligned} \tag{3.12}$$

Network Architecture

Network architecture is the blueprint of building the network. In this case, this shows how many hidden layers are present between the input and output layer and how many hidden units are present in each layer. According to the Universal Approximation Theorem , a feedforward network having a minimum of one hidden layer with any activation function is able to generate approxmation for any Borel observable function if we can provide enough hidden units for the network. To be simple, there is an MLP that can describe our desire function, whatever it it.

Well, we know that there might exist one or more MLP which can solve our problem but there is no standard method. We don't know how much layers and units are needed, so hit and trial method is the only way to determine this yet.

Finding this architecture is quite a challenge as we might need to consider different approaches and even we find the most suitable MLP architecture, it still might not be able to represent the goal function. Two reasons can be stated for this, first the optimization algorithm might perform poorly and so the values of the parameters might not be right and second, overfitting might occur.

Cost Function

Cost function is a function that determines how the network is as a whole. At any given point, this function shows the difference between the selected model's estimation and the goal. This function is single valued because it represents only one thing, the overall score of the network. If a feedforward neural network is trained using gradient based learning, stochastic gradient descent minimizes the cost function. The training phase heavily relies on the cost function so, choosing it needs proper research.

We can express cost function in this form C(W,B,Sr,Er).

Here W = weights of the nodes, B = biases of the network, Sr = Single training sample input, Er = Training sample desired output.

Now, to become a cost function, two properties must be satisfied:
1. The cost function needs to be expressed as an average.
2. Only activation values from the 3 layers can determine the function.

Mean Squared Error: In our model we have used MSE as cost function. MSE measures the average squared difference between the actual and predicted values of an observation. This function outputs the cost associated with the current set of weights. Less MSE mean more accurate model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{3.13}$$

Dropout

Dropout is a technique to prevent complex and powerful models of neural networks from overfitting. Feed forward neural network is more vulnerable to overfitting because of its high adaptivity. Dropout prevents this occurrence by temporarily dropping some of the units and their connections. This seems like crippling the network but making the network light reduces overfitting.

Dropout value for a node in a layer ranges from 0.0 (no output) to 1.0 (no dropout). The dropout value for our model is 0.5.

Verbose

Verbose is an option for producing detailed logging information. The higher the value is compared to 0, the slower the process is. Verbose for our model is 0.

Cross Validation

Cross-validation is the benchmark for a particular neural network. This helps to select the best set of parameter values from many.

Hyperparameter Tuning

Hyperparameters are model-specific properties that are set before the model is trained. Hyperparameter Tuning is the process of finding the right set of hyperparameter to achieve accuracy. There are two ways of hyperparameter tuning:

i) Grid Search: We test the model by trying every possible set of hyperparameters and evaluate the model for each combination. The pattern resembles a grid because for input we put all the values in matrix form. Set of parameters will be tested and remarked and finally the one with the top accuracy will be selected.

ii) Random Search: It is a strategy to figure out the finest solution using arbitrary combinations of hyperparameters. This attempts to combine a range of values at random. To optimise , the evaluation of the function is performed at arbitrary configurations in parameters space.

Hidden Units

Up next, we need to select the hidden unit. It can not be assured that any given unit is better than the other units so, choosing hidden units is also a critical task. However, there are some pre-selected units for beginners. The process of selecting a hidden unit needs to involve trial and error and use intuitive power.

Output units

These are the units in the output layer. These units give us the desired output, for which the neural network was designed in the first place. Output units are largely dependent on cost function. Hidden units can also be utilized as output units too.

3.6 Error Calculation

RMSE

The RMSE or the Root Mean Square Error is the square root of residuals or prediction errors. The RMSE value indicates how the predicted value performed comparing to the actual output. The distance between the regression line and the data points can be called as the residuals. RMSE shows the data cluster surrounding the best fit and shows the residuals distance from it. If considered for the measure of fit, the RMSE is absolute where R-squared is relative. The lower the value of RMSE the better the model predicts the output.Thus, the Root Mean Squared Error is an important estimation of the models forecasting capabilities, and if the model's main purpose is prediction, it is the most important criterion to verify.

R-Squared

The variability amongst SSE and SST is the change of the regression model's estimate over the mean model. It gives R-squared to divide the difference by SST. Relative to the mean model, it is the relative gain in estimation of the regression model. It indicates the model performance in terms of prediction.

The value for R-squared range from 0 to 1 to show prediction model's perfection. The value 0 indicates that the model is failed to predict the desired output wheres, value 1 shows that it perfectly predicts output for all data points.

Chapter 4

Market Analysis

4.1 Equity Indices

4.1.1 S&P500 (SP500)

S&P500 represents index of stock market that reflects stock price of top 500 companies in the US stock exchange. Any change in the US economy is directly linked to these companies. In recent times, we have seen that crude oil index and S&P500 is related in a strong positive manner. Two reasons can be linked with this. First, rise of S&P500 index means expected economic growth in the US. Increased economic activities mean increased demand for commodities which results in higher oil prices. Second, expansion of commodities indices has allowed oil to be a new investment which acts like stocks. This means the likelihood of both stock and oil prices move at the same pace.

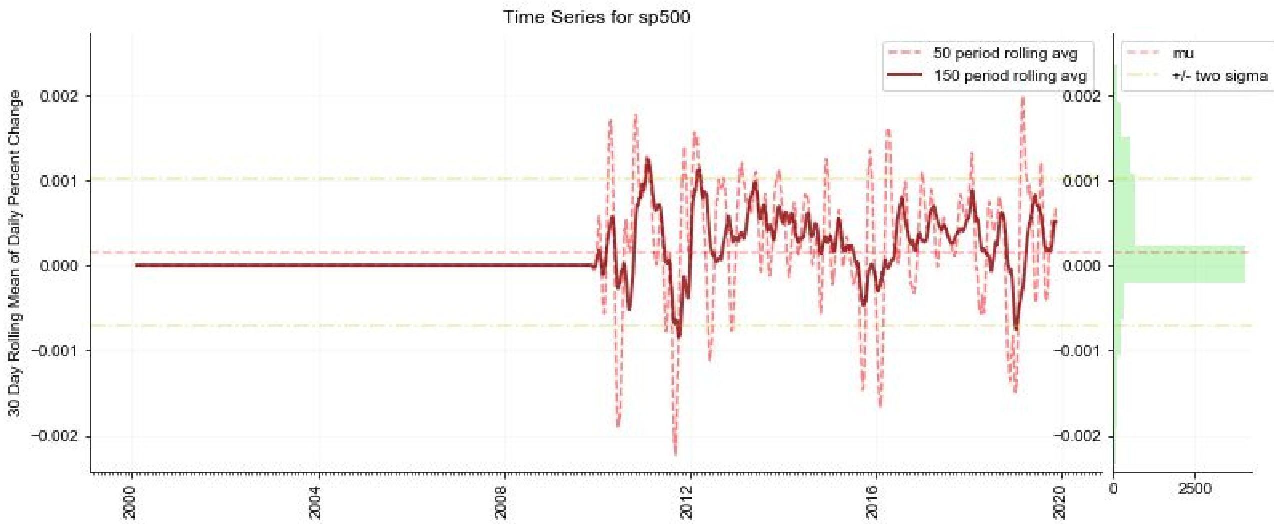


Figure 4.1: Percent Change of S&P500

4.1.2 NASDAQ Composite Index (NASDAQCOM)

The Nasdaq Stock Market, also called Nasdaq, is an American stock trade situated at One Liberty Plaza in New York City. It is positioned second on the rundown of stock trades by market capitalization of shares traded, behind just the New York Stock Exchange. Nasdaq, Inc. owns this platform.

The index which is used to measure this is the NASDAQ Composite Index, NASDAQCOM in short. This is the second ranked index in US Stock Exchange platform and correlates with the crude oil pricing just like S&P500.

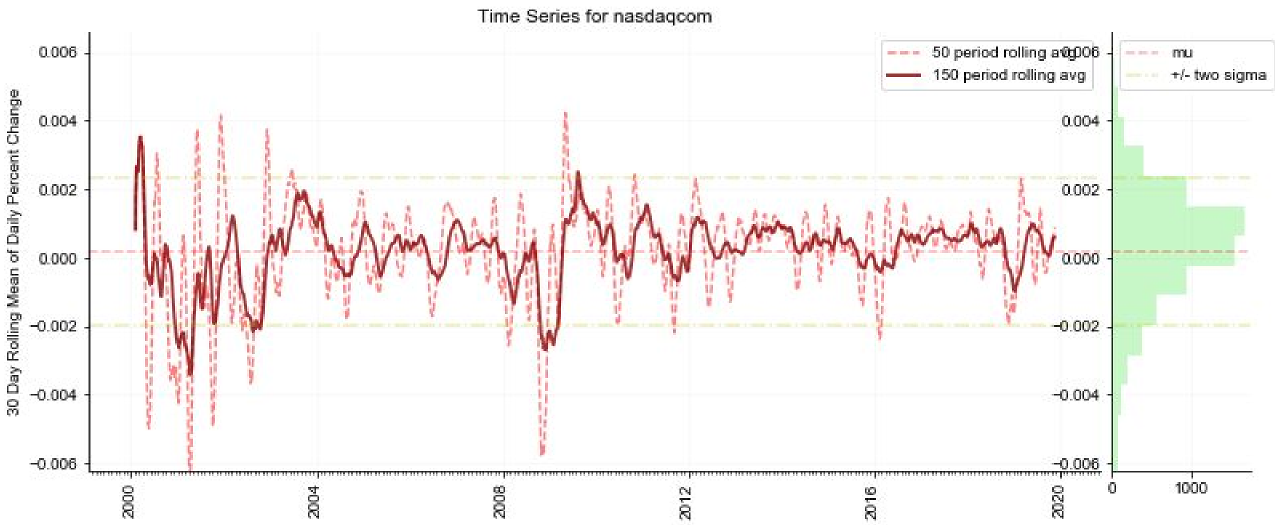


Figure 4.2: Percent Change of NASDAQ Composite Index

4.1.3 Dow Jones Industrial Average (DJIA)

Dow Jones Industrial Average is a stock market index that represents the stock value of the 30 large organizations recorded on the US stock exchange. Like S&P 500 this index also represents the U.S economy and so can be a good attribute for predicting crude oil price.

In spite of the fact that it is one of the most generally pursued stock indices, since it just incorporates 30 organizations and isn't weighted by market capitalization and weighted mean, many consider the Dow to not be a decent portrayal of the U.S. stock exchange and suggests that the S&P 500 is more reliable as it also includes the 30 companies of the Dow and 470 more. However, the Dow index still plays a vital role in controlling crude oil prices because of its huge acceptance as an indicator of the stock market of United States.

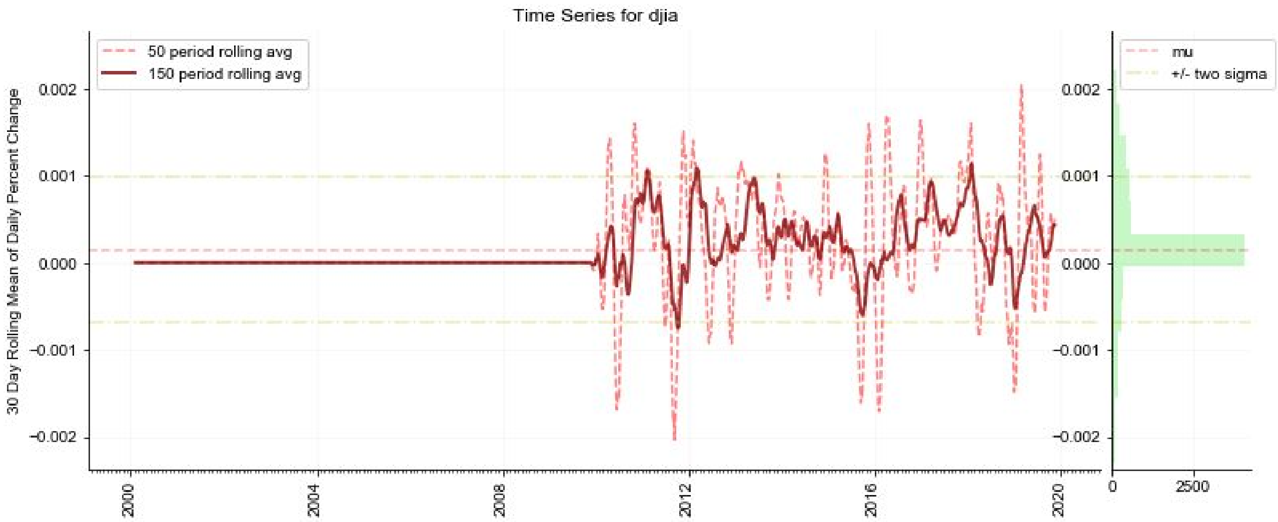


Figure 4.3: Percent Change of Dow Jones Industrial Average

4.1.4 Nikkei Stock Average, Nikkei 225 (NIKKEI225)

Nikkei 225 is a stock market index that represents the Nikkei Stock Average which consists of the top 225 stocks of the Tokyo Stock Index based on liquidity. This index shows the daily index value at market close.

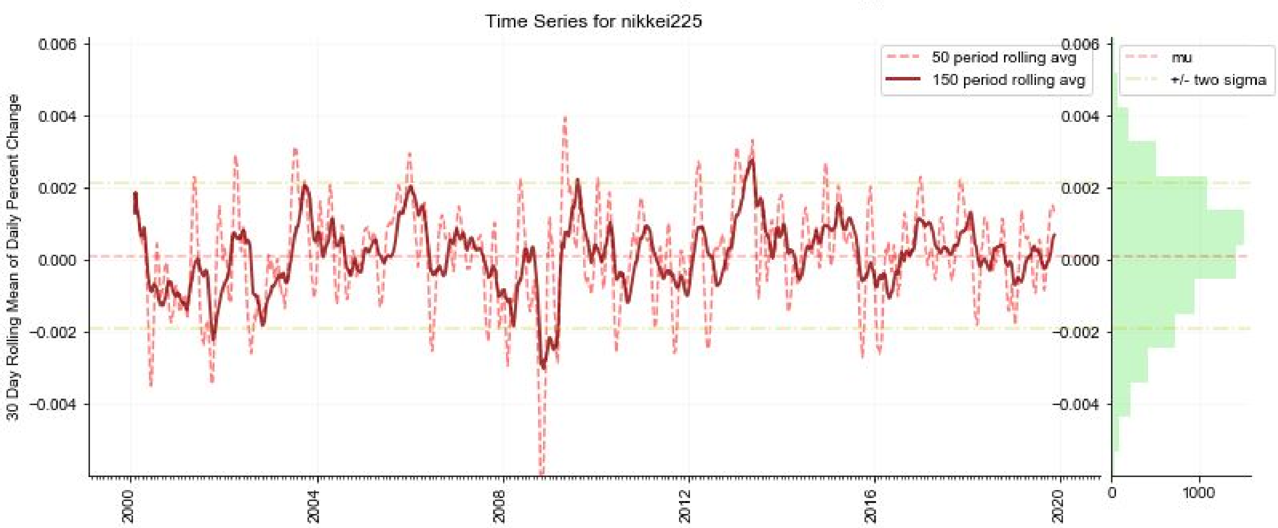


Figure 4.4: Percent Change of Nikkei Stock Average, Nikkei 225

4.1.5 Interest Rate on Excess Reserves (IOER)

Excess reserves are the amount of excess reserve compared to the standard measure. The Board of Governors considers the amount and put an interest rate on this excess reserve. The Federal Reserve uses this measure as an tool to conduct monetary policy suitably.

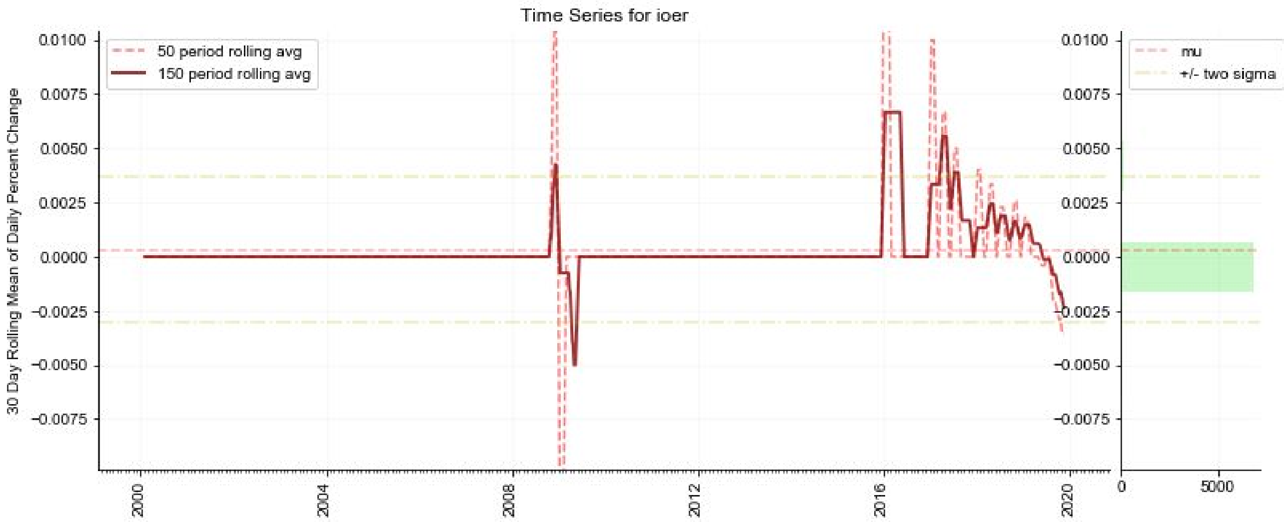


Figure 4.5: Percent Change of Interest Rate on Excess Reserves (IOER)

4.1.6 CBOE Volatility Index (VXXLECLS)

Volatility Index, commonly known as VIX is a real-time index that is derived from the price inputs of S&P500. This index is generated by CBOE which is the acronym for Chicago Board Option Exchange which predicts the expected market volatility for the next month. Thus, it gives the general overview of risk for the investors which is why this index is also known as the "Fear Index". As VIX values can be a way to measure the economy, it also plays an important role in measuring the change in oil prices.

4.2 Curency Exchange Rate

4.2.1 Monetary Base (BOGMBASEW)

The monetary base of the US, which means the total amount of bank notes circulating in the country. The total amount of currency in the hands of the public, currency that is physically held in the vault of the banks and the commercial bank reserves that is held in the central bank.

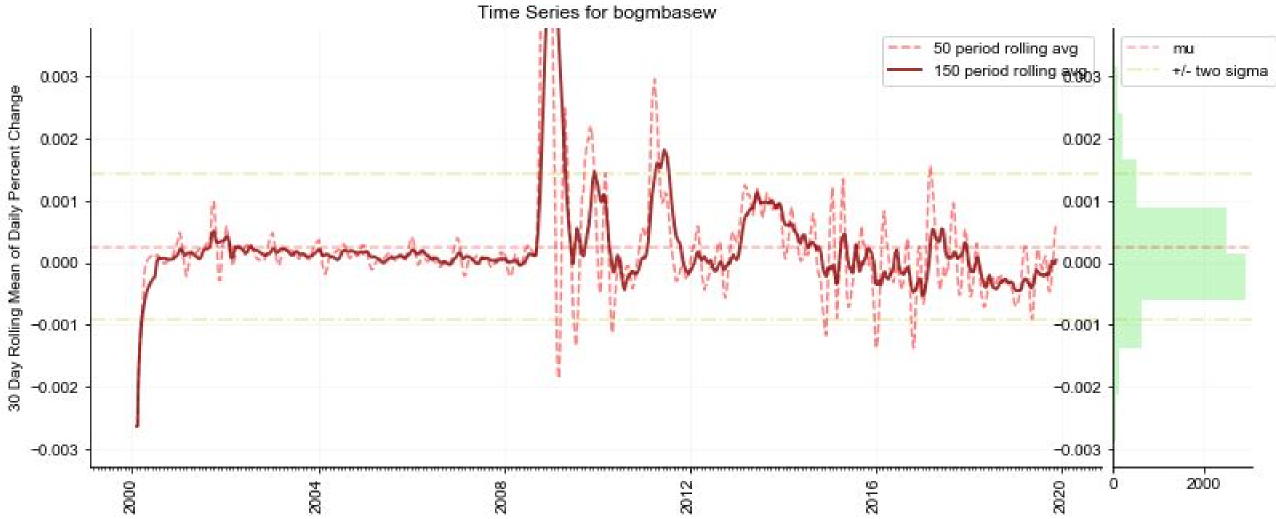


Figure 4.6: Percent Change of Monetary Base

4.2.2 DEXJPUS

Japanese Yen to One U.S. Dollar, Not Seasonally Adjusted

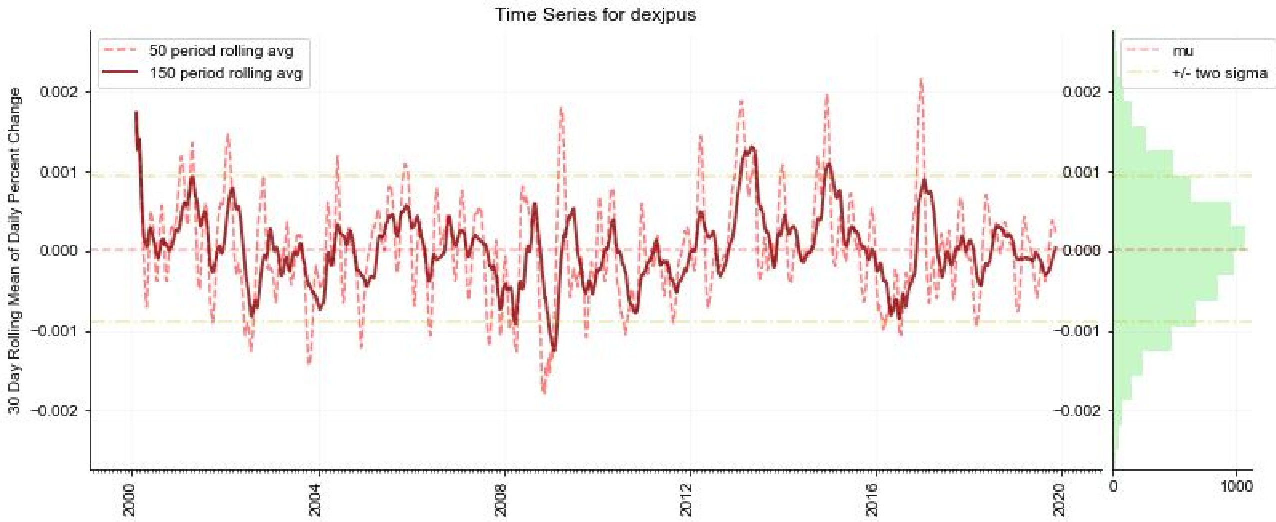


Figure 4.7: Percent Change of Japanese Yen to One U.S. Dollar

4.2.3 DEXUSEU

United States Dollars to one Euro which is not modified seasonally.

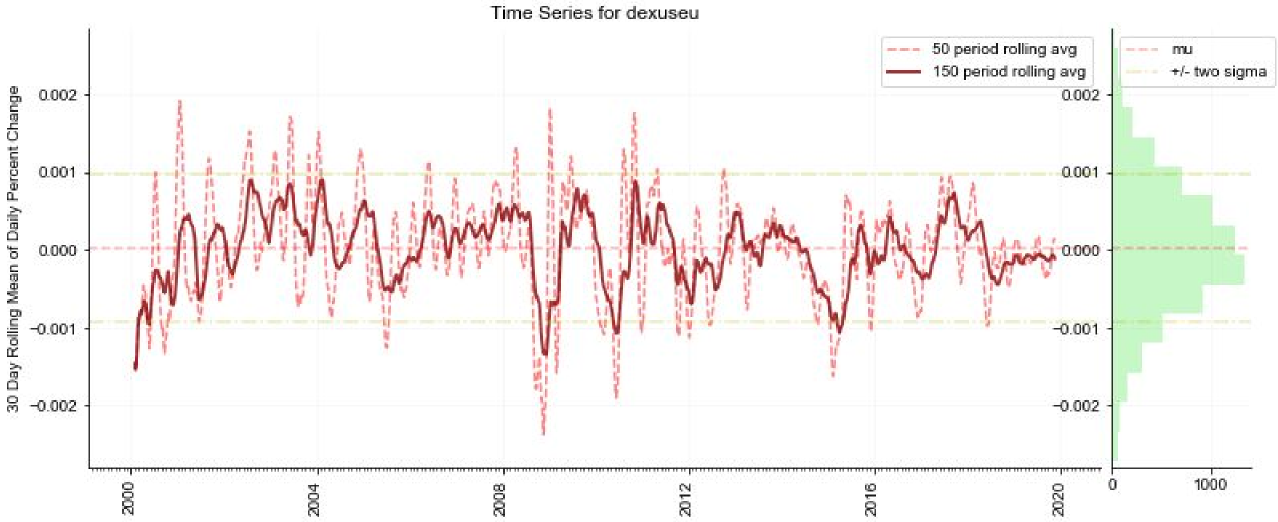


Figure 4.8: Percent Change of USD to one Euro

4.2.4 DEXCHUS

CNY to one USD. Daily, Not Seasonally Adjusted

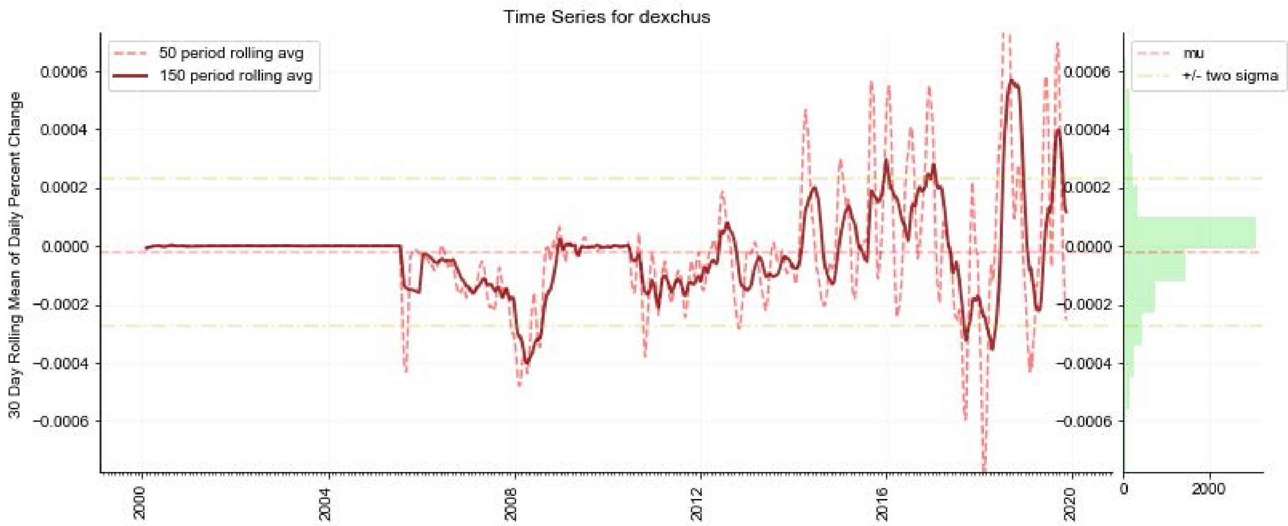


Figure 4.9: Percent Change of CNY to one USD

4.2.5 DEXUSAL

USD to One Australian Dollar, Daily and not modified seasonally.

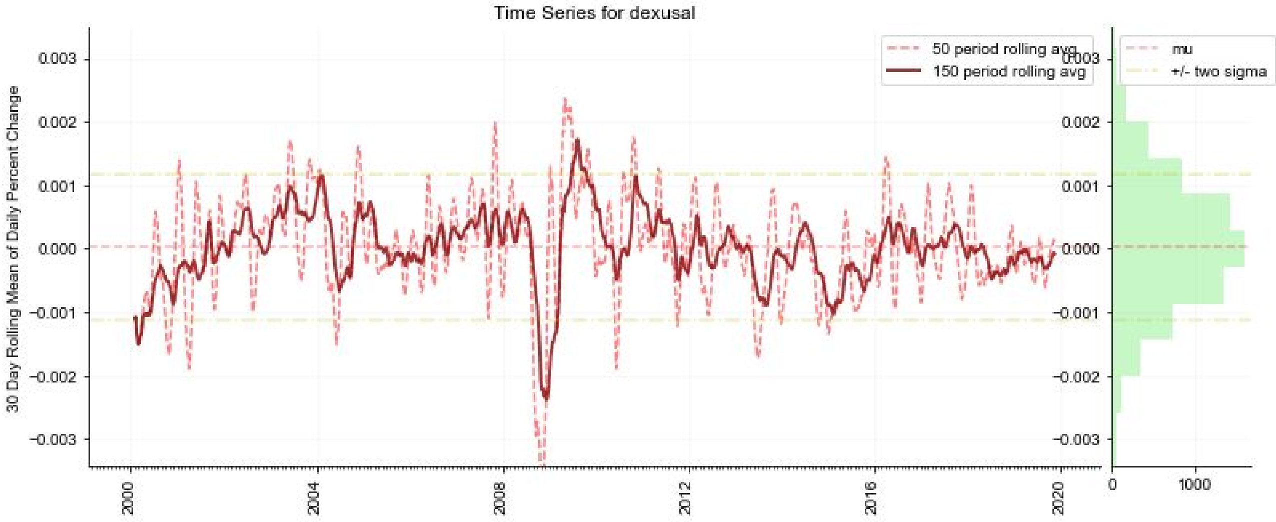


Figure 4.10: Percent Change of USD to one AUD

4.3 Debt Market Indicator

4.3.1 USDONTD156N

It is the average of loan fee at which top banks take loans from other banks. This indicator as also knows as LIBOR (London Interbank Offer Rate) which is the most generally utilized "benchmark" or reference rate for momentary loan fees.

As the LIBOR contributors and ICE Benchmark Administration Limited (IBA) planned, the endorser states that, to the furthest reaches allowed by regulation, the LIBOR contributors and the IBA are not responsible for any risk of predicting or determining the rate regardless of whatever the reason is and must bear the risk of any kind of loss or harm because of any inconsistency or error.

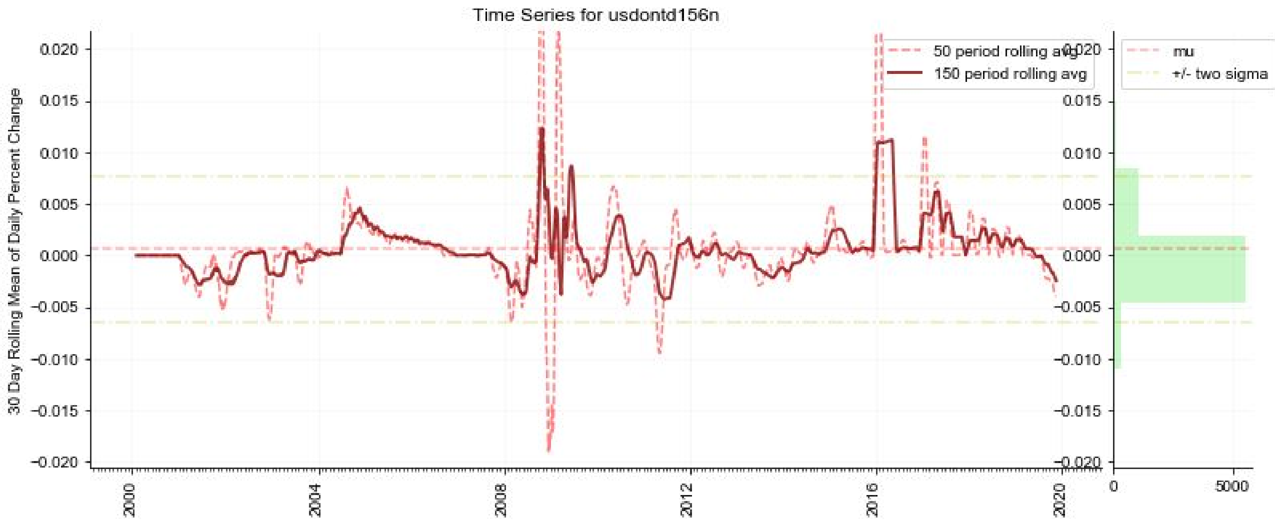


Figure 4.11: Percent Change of London Interbank Offered Rate Average Loan Fee

4.3.2 USD1MTD156N

This is the monthly London Interbank Offer Rate (LIBOR). All preconditions and variables are same as LIBOR which is already explained.

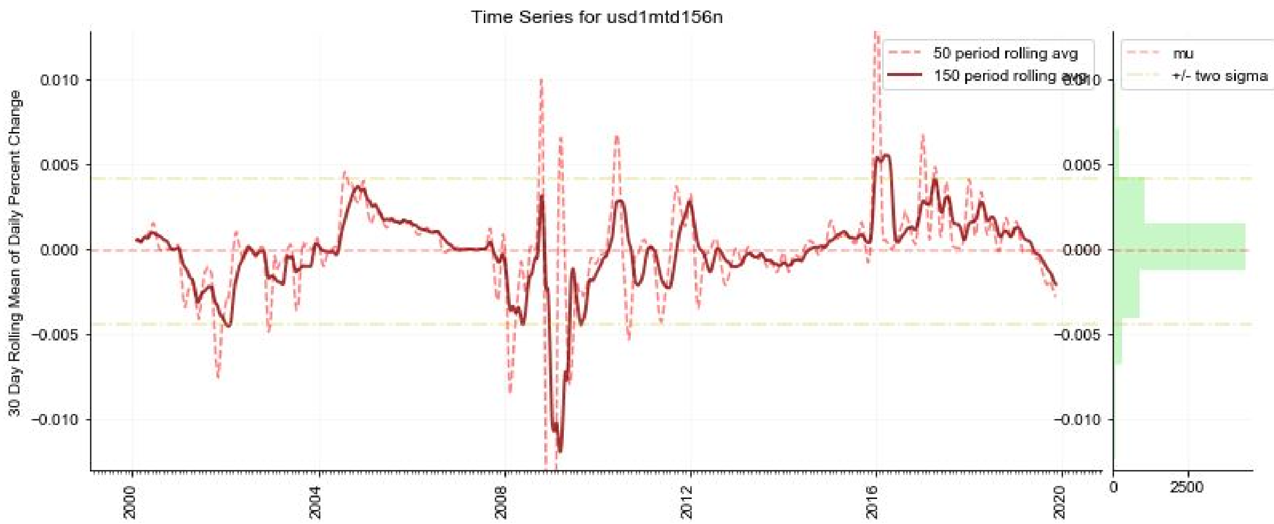


Figure 4.12: Percent change of LIBOR in one month

4.3.3 USD3MTD156N

This is the LIBOR for the 3 months. All the preconditions and variables are same as LIBOR which is already explained.

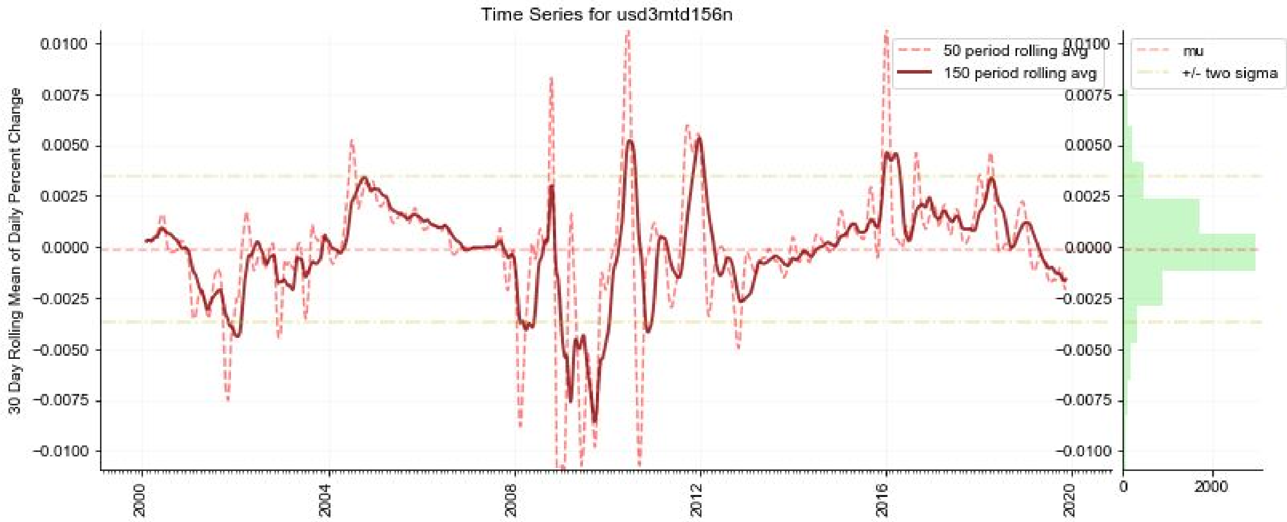


Figure 4.13: Percent change of LIBOR for 3 months

4.3.4 USD12MD156N

This is the yearly LIBOR. All the preconditions and variables are same as LIBOR which is already explained.

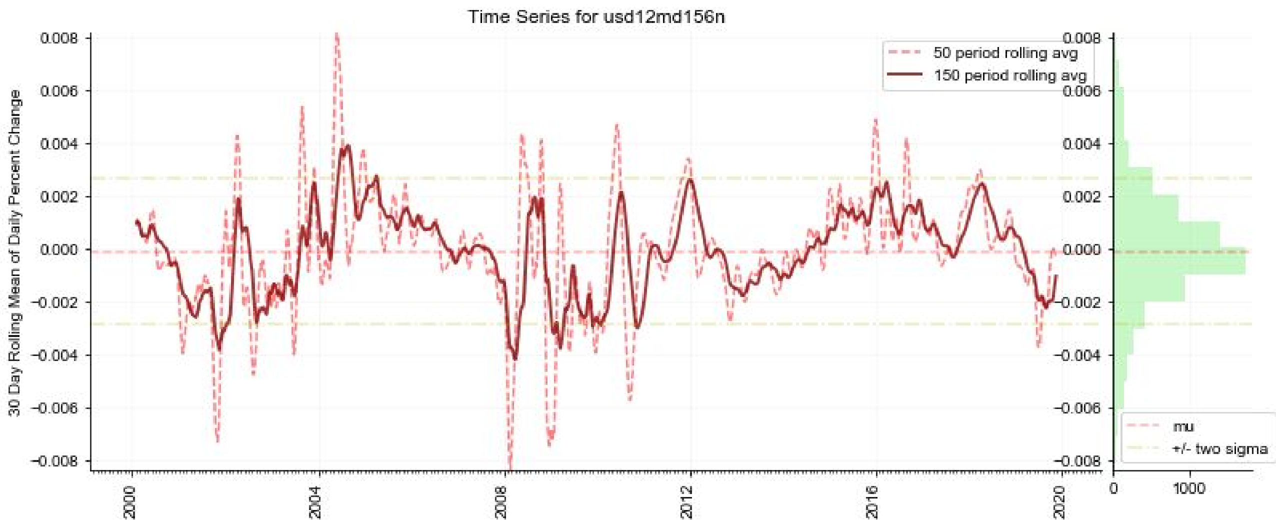


Figure 4.14: Yearly percent change of LIBOR

4.3.5 BAMLHYH0A0HYM2TRIV

This attribute represents the dollar-denominated corporate debt when dollar bonds are issued publically in the US domestic market.

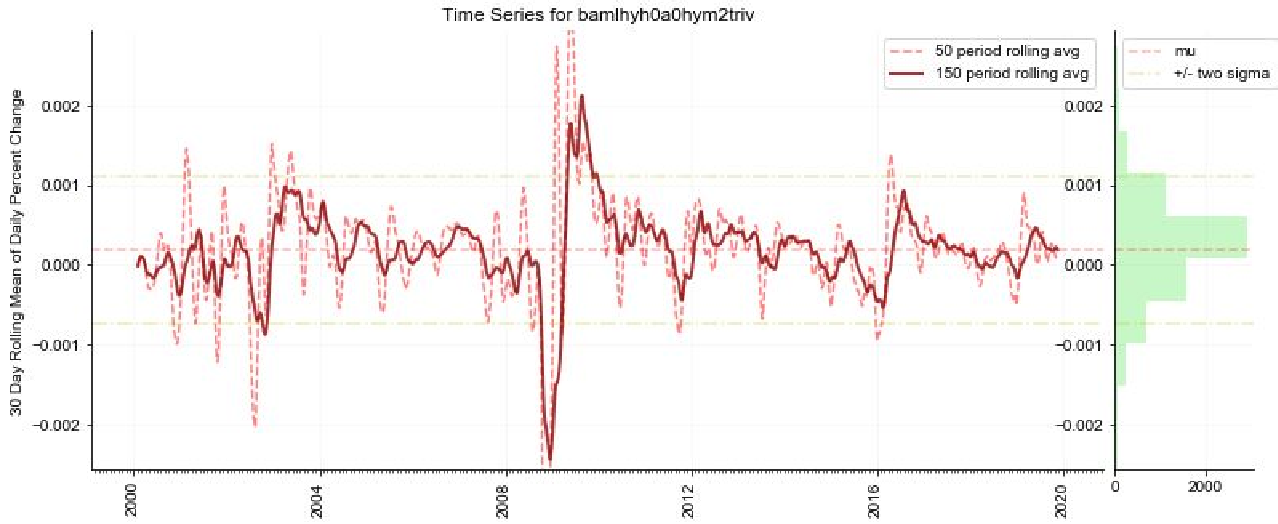


Figure 4.15: Percent Change of ICE BofAML US High Yield Master II Index

4.3.6 ICE BofAML US Corp AAA Total Return Index Value (BAMLCC0A1AAATRIV)

This index tracks the corporate debt that was issued to the public in the US market. This incorporates the securities with a provided speculation grade rating AAA. If the last day of the month is a weekend, the output will be a representation of increasing interest adjustments.

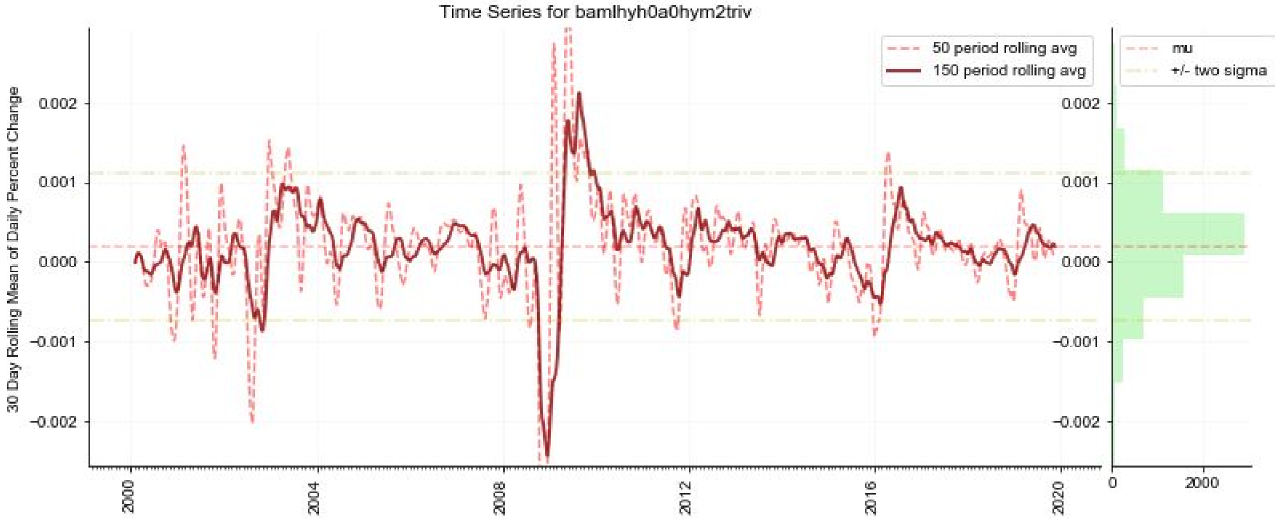


Figure 4.16: Percent Change of ICE BofAML US Corp AAA Total Return Index Value

4.3.7 ICE BofAML Euro High Yield Index Option-Adjusted Spread (BAMLHE00EHYIOAS)

This is the measure of euro-denominated corporate debt when euro bonds are issued publicly in the euro domestic market.

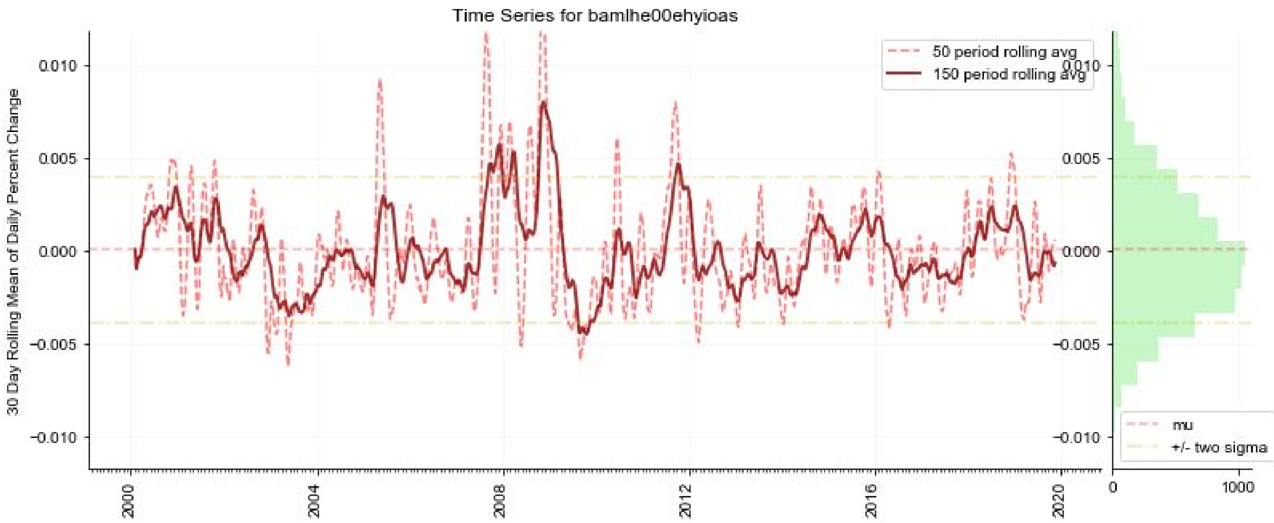


Figure 4.17: Percent Change of ICE BofAML Euro High Yield Index Option-Adjusted Spread

4.4 Comodity Prices

4.4.1 Gold Fixing Price (GOLDAMGBD228NLBM)

The London Bullion Market Association Gold Price was propelled on the 20-03-2015 to supplant the notable London Gold Fix. LBMA holds the intellectual rights but the auction stage, methodology and command for the LBMA Gold Price which is set by the ICE Benchmark Administration (IBA). The value keeps on being set at 10:30 once and 15:30 twice everyday on London Time in USD. Intead of U.S. Dollars, Euro and Sterling prices are taken only to be used as an indicative price.

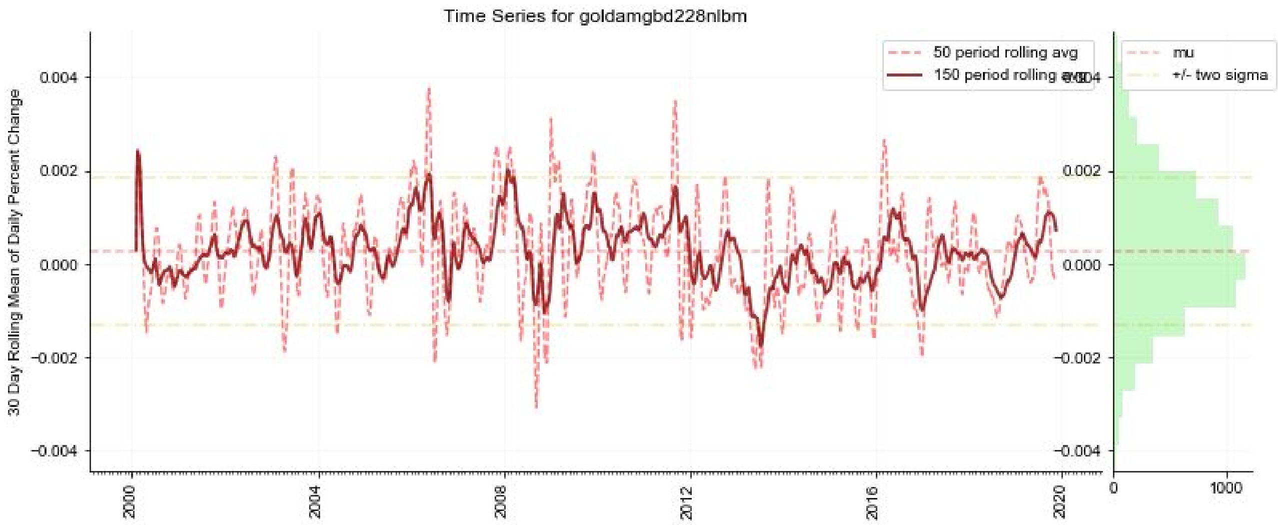


Figure 4.18: Percent Change of Gold Price

4.4.2 West Texas Intermediate (WTI) (DCOILWTICO)

For a long time, WTI Crude Oil has been utilized as a standard in crude oil pricing because of its high acceptance rate throughout the world.

Range is 1986 to 2019. Unit is dollars per barrel. Frequency is daily. The data is not seasonally adjusted.

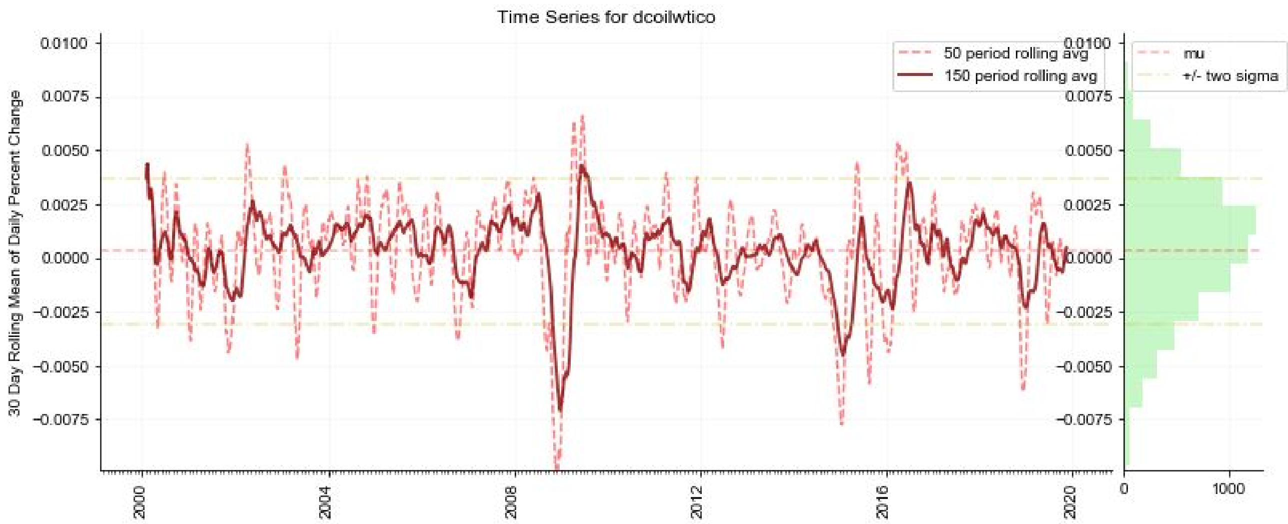


Figure 4.19: Percent Change of WTI Crude Oil Prices

4.5 Energy-Related Series

4.5.1 Henry Hub Natural Gas Spot Price (MHHNGSP)

It is the indicator of the amount of US Dollar needed to purchase one million BTU natural gas.

Range is 1997 to 2019. Unit is dollars per million BTU. Frequency is daily. The data is not seasonally adjusted.

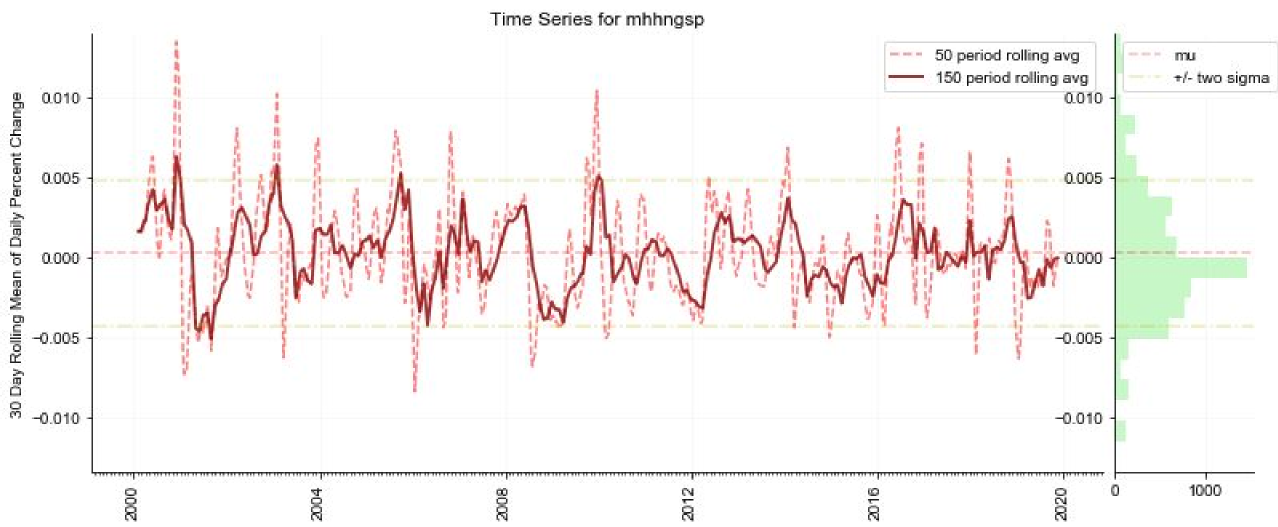


Figure 4.20: Percent Change of Henry Hub Natural Gas Spot Price

4.5.2 CBOE Energy Sector ETF Volatility Index (VXXLECLS)

ETFs are trusts shares that monitor the price index and other indices to analyse the market performance.

4.6 Geopolitical Factors

4.6.1 Equity Market Volatility Tracker: Elections And Political Governance (EMVELECTGOVRN)

The Equity Market Volatility tracker moves with the VIX and with the realized volatility of returns on the SP500.

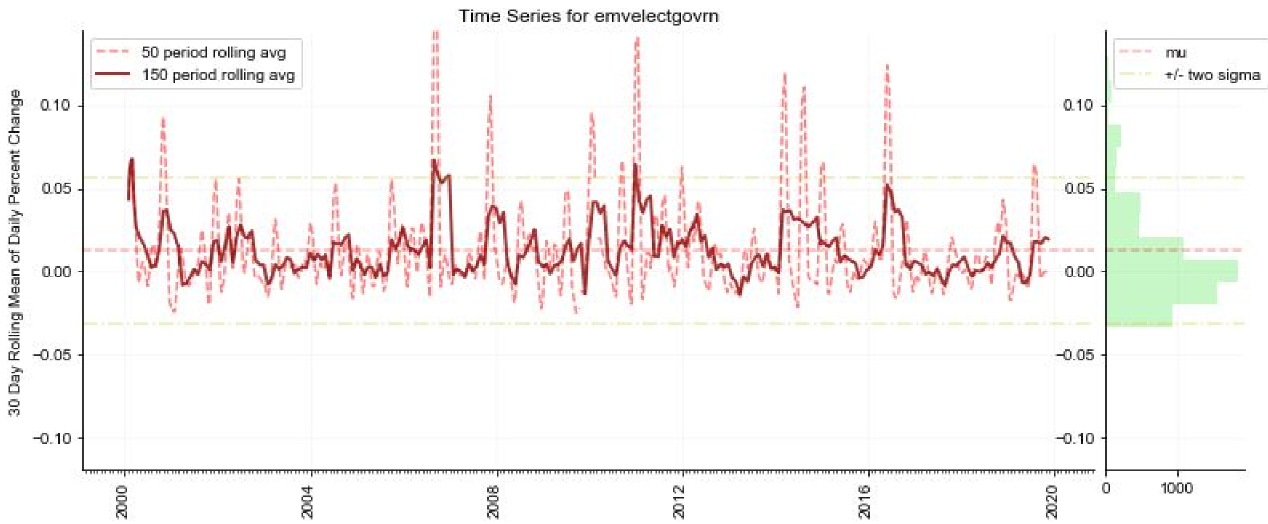


Figure 4.21: Percent Change of Equity Market Volatility Tracker: Elections And Political Governance

4.6.2 All Employees: Business, Professional, Labor, Political, and Similar Organizations in New York (SMU36000008081390001SA)

This sequence is seasonally adjusted by the St. Louis Federal Reserve Bank using Python's' statsmodel' module with default parameter settings. Using the U.S. kit. Seasonal Adjustment System Office of the Census X-13ARIMA-SEATS. Further information on the X-13ARIMA-SEATS ' statsmodel' kit is available here. You can find more information about X-13ARIMA-SEATS here.

Many series contain the non-seasonally (NSA) as well as seasonally adjusted data (SA). NSA data will be updated in some situations, but the SA data won't be changed. The explanation is generally that not enough new seasonal variables have been collected by the data series to cause a change. The NSA series can be found here The FRED team is currently working on a new procedure to replace data from SA that has not yet been updated with updated NSA data.

Many seasonally adjusted series may have negative values because they are created from a system of seasonal adjustment regardless of the actual meaning or perception of the indicator.

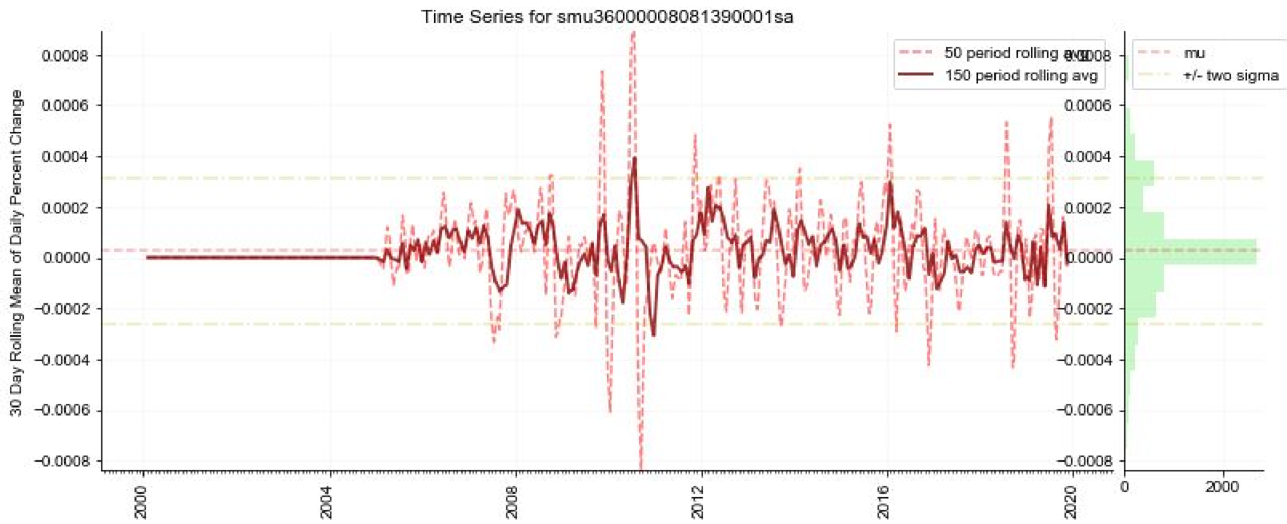


Figure 4.22: Percent Change of All Employees: Business, Professional, Labor, Political, and Similar Organizations in New York

4.7 Country Wise Macroeconomic Factors

We used two datasets to predict the oil price. One of these contains the macroeconomic factors of U.S and the geopolitical risk factors and the other one contains macroeconomic factors for the top oil using countries because their interest in buying crude oil affects the price largely. The attributes that were used to train the model are described here.

4.7.1 GDP growth

GDP or Gross Domestic Product is the aggregate value in money for all products that are being produced within a country in a given year. This indicates the economic condition of a country.

GDP is used to measure the countries economic development through comparing the GDP for one quarter with another quarter. In terms of oil, GDP and oil production had strong positive correlation. With the increase of oil production the GDP increases as well because oil is one of the prime fuel for production in any country.

4.7.2 CPI

A systematic calculation used to estimate changes of price in a basket of products that represents the expense of consumption is called Consumer Price Index of that specific economy.

The analysis used in CPI's estimate is very rigorous. For the classification of consumer items, various categories have been developed. The final overall price index is determined by national statistical firms based on these indices collected. CPI is one of the most important economic indicators which is depended on weighted average of the product prices which gives the perception of the living condition of the citizens of a country.

CPI is used to calculate inflation. Over a given period, the percentage increase in CPI is being calculated to find out the inflation of a country. Through the increase rate of CPI it can be called that the people need to pay more than what they used to with the same amount of money.

CPI and oil price is directly connected. The commodity prices go up with the increase of oil price as oil is used for production activity in most countries.

4.7.3 Interest Rate

The amount of money that is charged by the bank for providing the loan. Usually this is expressed as an annual percentage rate (APR) of the loan.

To put simply, interest is the money that a borrower need to pay for using any assets. In terms of big property such as, a house or vehicle the lease rate the borrower give can also be considered as interest rate. The interest rate is depend on the risk level of the borrower in terms of returning the property. If the borrower has high risk the interest risk is also high and it becomes lower otherwise.

The interest rate on loans is added to the balance, which is the loan volume. The interest rate is the borrower's loan price and the lender's return level.

4.7.4 Production

Production indicates crude oil production in the specified country in a given year. Unit of measurement for crude oil is BBL which is 42 U.S. gallons. Oil production and price has inverse relationship. With the production boom the supply of oil increases thus the price drops down.

4.7.5 Consumption

Consumption indicates the consumption of crude oil in the specified country in a year. The unit of measurement for crude oil is BBL which is 42 U.S. gallons. Consumption gives the demand curve for a given product. If the consumption rate goes high and the supply of oil does not meet the demand then the price of oil increase accordingly.

4.7.6 Oil Rent

Generally, the term oil lease is used to refer to the discrepancy between crude oil's selling price and its cost of production. Therefore, it is the exploration and production activity's income before tax. The taxes paid to the state where production happens are part of the oil lease share of the state.

Renting oil is a rent for Ricardians. This occurs when a product's sales price is higher than the least efficient producer's marginal cost.

4.7.7 Employment

Employment rates are characterized as a measure of the use of available labor capital (people available for work). The employment rate is calculated by the proportion between the population that are looking for jobs and the population that are employed in a given time. Criteria for employment varies but employment rate always indicates the allocation of resources and workplace for its population. The higher it is, the better the country's economy is because the country can use its resources with more efficiency.

4.7.8 Military Spending

The percentage of a country's annual budget that is spent on the military. Usually, military spending is quite high for any developed country so, this attribute is a vital one for understanding a country's economy and financial decision making.

4.7.9 GPR

GPR stands for geopolitical risk. Geopolitics, regional impact research of international relations on power relations. Geopolitics is described as the study of the relationship between geographical settings and viewpoints, on the one hand, and political processes, on the other.

GPR tries to identify strategic threat where the power disputes of agents over territory can not be resolved peacefully or legitimately. We therefore describe geopolitical risk as the risk of conflicts, terrorist attacks, disputes between states hindering the international tranquility among the nations. Geopolitical risk encompasses both the danger of the continuation of these incidents and the additional risks associated with the worsening of current events.

The GPR index is being constructed by computing the number of incidents in the top newspapers from 1985 till present showing geopolitical events and risks using automated text and heading searches from the electronic archives. We gathered the data from the following 11 newspapers - The Daily Telegraph, The Guardian, The Boston Globe, The Globe and Mail, The Chicago Tribune, The Times, The New York Times, The Los Angeles Times, The Washington Post, The Wall Street Journal and The Financial Times.

Chapter 5

Data Processing

5.1 ARIMA Model

In order to train the ARIMA model we used only the daily WTI crude oil price as the prediction factor. We generate the dataset from "quandl" API. Quandl provides financial and economic dataset for stock market, oil market and so on. We called the daily crude oil price from January 2003 until November 2019 by API of quandl through –

```
quandl.ApiConfig.api_key = 'xszj_ShXe7_Z6hgGpWd2'
oil = quandl.get('OPEC/ORB', start_date='2003-01-01', end_date='2019-12-01') # This creates pandas data frame
```

Figure 5.1: Fetching Data Using Quandl API

Afterwards, to justify if the data is imported correctly both the first and last five values of the imported dataset was monitored –

Date	Value
2003-01-02	30.05
2003-01-03	30.83
2003-01-06	30.71
2003-01-07	29.72
2003-01-08	28.86

Date	Value
2019-11-25	64.21
2019-11-26	63.92
2019-11-27	64.40
2019-11-28	63.94
2019-11-29	63.83

Figure 5.2: First Five Values and Last Five Values

In order to work with time series it is cumbersome to use the string value of date and time to predict future behaviour. Thus the values were being converted to Julian date to work with easily.

5.1.1 Julian Date

JD which is the acronym for Julian date is represented by integer and fraction number where the count of days since initial epoch and the fraction represents the time of day [26]. Julian date is widely used in prediction analysis as it gives a continuous floating value of different days and time.

After converting the dataset into Julian the first and last five values obtained are –

Date	const	Value	Julian
2003-01-02	1.0	30.05	2452641.5
2003-01-03	1.0	30.83	2452642.5
2003-01-06	1.0	30.71	2452645.5
2003-01-07	1.0	29.72	2452646.5
2003-01-08	1.0	28.86	2452647.5

Date	const	Value	Julian
2019-11-25	1.0	64.21	2458812.5
2019-11-26	1.0	63.92	2458813.5
2019-11-27	1.0	64.40	2458814.5
2019-11-28	1.0	63.94	2458815.5
2019-11-29	1.0	63.83	2458816.5

Figure 5.3: First Five Julian Values and Last Five Julian Values

We had gathered a total of 4364 data to taste and train data to predict oil price –

```
oil.shape
```

```
(4364, 3)
```

Figure 5.4: Dataset shape

We plotted the price of oil according to the conventional calendar to show the changes of price with time and volatility associated with it –

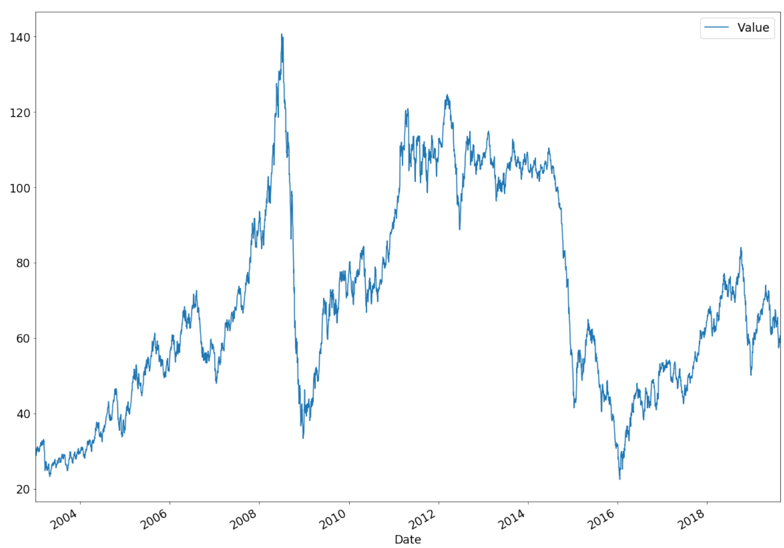


Figure 5.5: Oil price over time

In time series analysis we need to consider the followings conditions:

5.1.2 Stationary

The term stationary in time series signifies that the statistical value (e.g. mean, variance, auto correlation) do not change over time. In other words it implies whether a time series has stationary characteristics if all $k \geq 0$ and t the following two k -tuples have the same distribution:

$$(A0, \ldots,) \sim (Xt, \ldots, Xt + k)$$

It is important to convert the series to stationary from non-stationary because it makes the time series easier to analyse.

5.1.3 Drift

In time series the term drift refers to the intercept component of a series. When a time series is non stationary it possesses drift characteristic. Mathematically it can be expressed as –

$$Xt = \mu t + \varepsilon t \tag{5.1}$$

5.1.4 Seasonality

To deal with time series it is mandatory to meet with the term seasonality. Seasonality in time series implies the variations that can be seen in different season of the year [3]. For example, sell of ice cream increases in summer and decreases in winter due to the temperature changes during the seasons. Mathematical representation of the seasonality might be –

$$Xt = a \sin(\omega t) + \beta \cos(\omega t) \tag{5.2}$$

5.1.5 Trend

Trend in time series is the upward or downward movement which happens in a long term and indicates the pattern that the time series likely to follow.

5.1.6 Residuals

Residuals are the components that left alone when calculating the value of other components. For the simplicity of calculation the residuals are being removed from the time series analysis.

The Dickey-Fuller test was performed to determine whether the obtained time series is stationary or not.

5.1.7 Dickey-Fuller test

The result from Dickey-Fuller test has the following indicators such as test statistics, critical values, lags used, p value and number of observations. Among them the test statistics and critical values are the determinant of the time series stationary properties. If the critical values are less than the test statistics then the null hypothesis gets accepted otherwise it is rejected.

5.1.8 Null Hypothesis (H0)

Whenever a null hypothesis gets accepted it shows the time series is non-stationary with unit root.

5.1.9 Alternate Hypothesis (H1)

We move to alternate hypothesis only if the null hypothesis gets rejected or the critical values are greater than the test statistics. It shows that the time series is stationary.

The test verifies if $\phi = 0$ in time series:

$$yt = \alpha + \beta t + \phi yt - 1 + et \tag{5.3}$$

That can be expressed as follows:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \gamma yt - 1 + et \tag{5.4}$$

In our observation the result from Dickey-Fuller test came as follows:

```
Results of Dickey-Fuller Test:
Test Statistic                   -2.328617
p-value                           0.162905
#Lags Used                       25.000000
Number of Observations Used    4332.000000
Critical Value (1%)              -3.431860
Critical Value (5%)              -2.862207
Critical Value (10%)             -2.567125
dtype: float64
```

Figure 5.6: Dickey-Fuller test result

From the Dickey-Fuller test result it was showed that the value of test statistic is significantly larger than the critical values. Thus our generated time series has non-stationary characteristics.

After plotting values for Original, Trend, Seasonality and Residual points we get the following graphs –

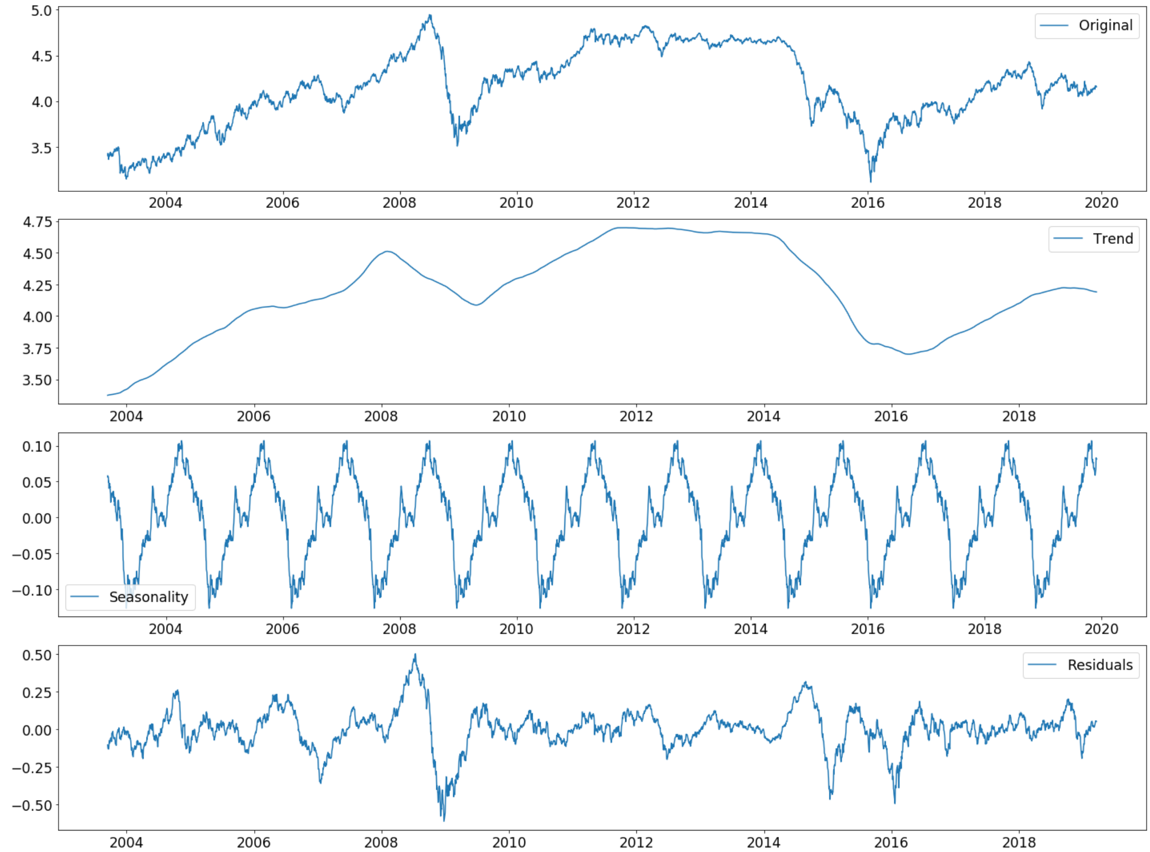


Figure 5.7: Graphs of Original time series, Trend, Seasonality and Residuals

The graph for trend showed that the time series did not show any trend pattern to follow. From the seasonality graph we could see that it did not have any seasonal effect. Lastly, the residual graph shows a significant number of residuals.

We used the residuals to perform a Dickey-Fuller test and also plotted the rolling mean and standard deviation curve using it. Which showed a relatively better outcome to work with:

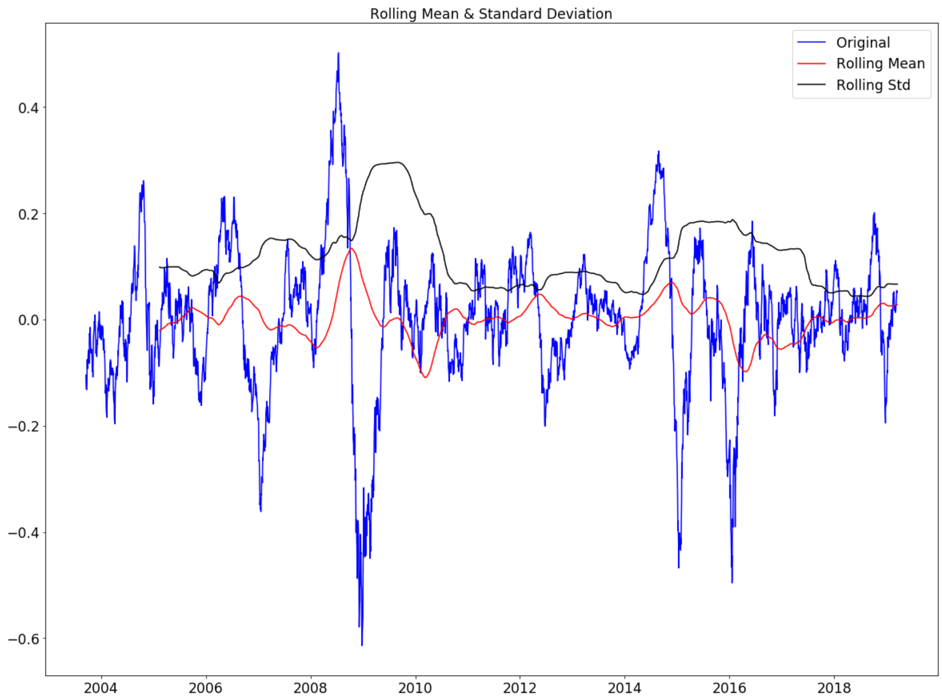


Figure 5.8: Standard Deviation and Rolling mean

And the test result –

```
Results of Dickey-Fuller Test:
Test Statistic                 -5.932488e+00
p-value                         2.362611e-07
#Lags Used                      2.500000e+01
Number of Observations Used     3.974000e+03
Critical Value (1%)            -3.431997e+00
Critical Value (5%)            -2.862268e+00
Critical Value (10%)           -2.567157e+00
dtype: float64
```

Figure 5.9: Dickey-Fuller test after using residuals

With those results we came to the decision for the value of the parameter d should be 1. To determine the value for the parameter q and p we plotted the autocorrelation and partial autocorrelation graph respectively.

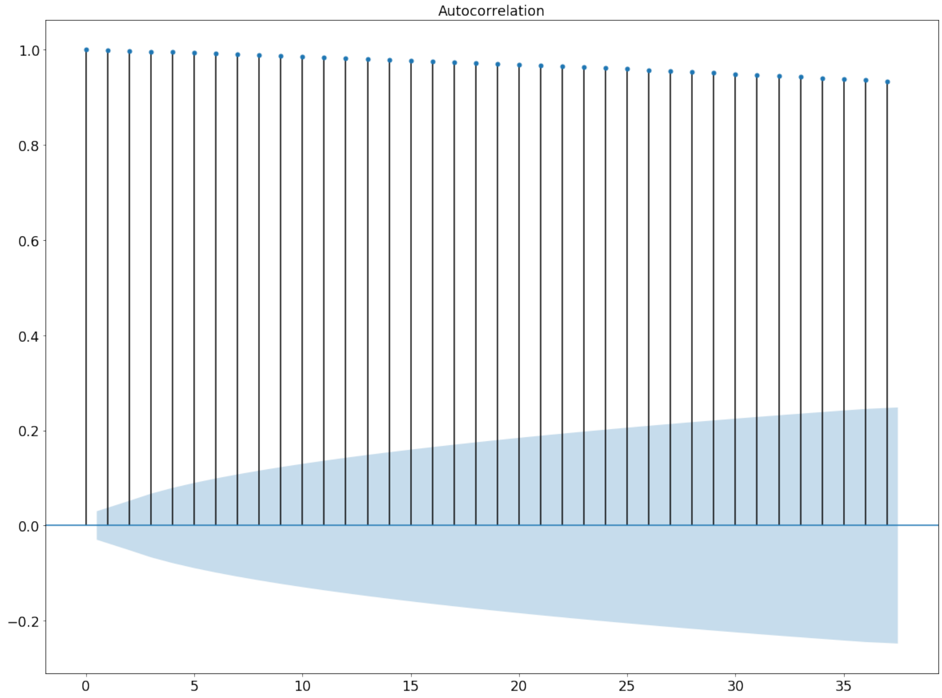


Figure 5.10: Graphical representation of Autocorrelation Function

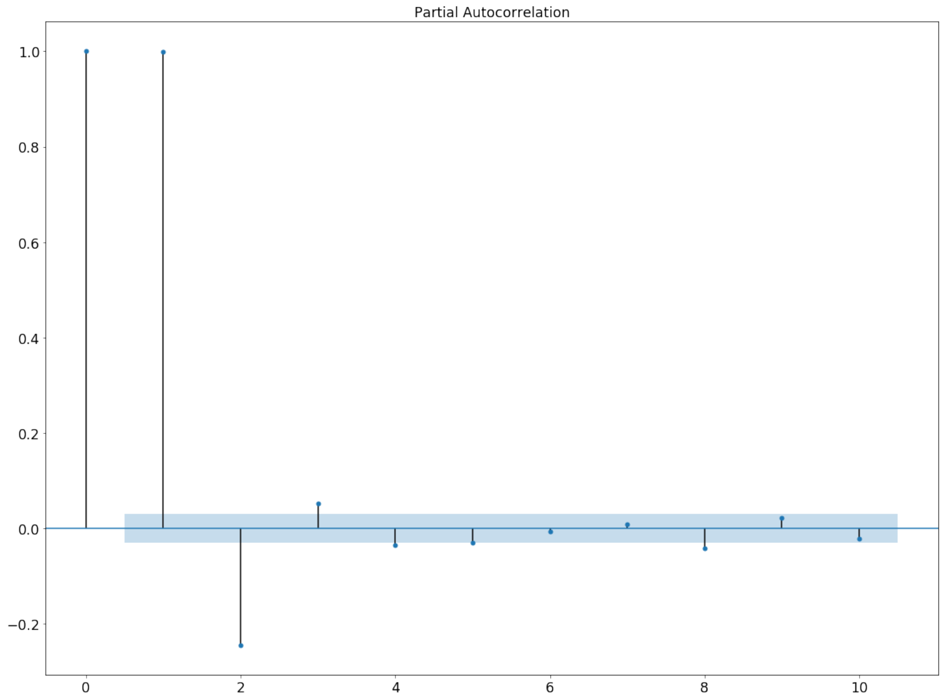


Figure 5.11: Graphical representation of Partial Autocorrelation

The autocorrelation graph showed that it moved to lower level for the first time at 2 so the value for q was determined to be 2. Similarly, the value for p was found as 1 from the partial autocorrelation graph.

Finally, we trained the model over 3500 data and test on 863 data and got the prediction of crude oil price for ARIMA model.

5.2 Feedforward Neural Network

We collected Macroeconomic data from FRED which is the acronym for Federal Reserve Economic Data. We called a function - get_fred_data which is defined to call the Saint Louis FRED API via pandas_datareader. From January 01, 2000 to 22 November, 2019 we fetched 7000 data and 24 factors for our prediction.

5.2.1 Generate Calendar

Python has an inbuilt calendar module which handles operations related to calendar. The calendar is Gregorian calendar which start the weekday from Monday and assume Sunday as the last day of the week. We used the calendar to divide the day into holiday and weekday to show oil trade on the particular day of the week.

date	month	year	weekday	is_weekday	is_holiday	is_holiday_week
2000-01-01	January	2000	Saturday	0	0	0
2000-01-02	January	2000	Sunday	0	0	0
2000-01-03	January	2000	Monday	1	0	0
2000-01-04	January	2000	Tuesday	1	0	0
2000-01-05	January	2000	Wednesday	1	0	0

Figure 5.12: Generate Calendar

5.2.2 Fill Missing Value

Pandas arrangement is a One dimensional ndarray with axis labels. The labels need not be exceptional however should be a hashable sort. The item underpins both whole number and name based ordering and gives a large group of techniques to performing activities including the record. Pandas Series.fillna() work is utilized to fill NaN esteems utilizing the predetermined technique.We used this technique to fill up the missing values that were not available via FRED API.

5.2.3 One Hot Encoding

One hot encoding is used for categorical data which do not contain integer or binary values but label values. By the one-hot encoding, each label values has unique binary value. So, One hot encoding is a procedure by which clear cut factors are changed over into a structure that could be given to Machine Learning algorithms to make a superior showing in forecast.

5.2.4 Rolling Mean Of Daily Percent Change

Percent change is a mathematical calculation which represents the change overtime. It is a continuous necessity in information investigation to figure how a variable has changed over a time span. A pandas DataFrame can be stacked with various time arrangement information of numerous factors, where every section of the DataFrame compares to a time series. The pct_change() strategy for DataFrame class in pandas processes the rate change between the rows of information.

$$percent\_change = [(final\_value - initial\_value) \div (initial\_value)] \times 100 \quad (5.5)$$

5.2.5 Correlation

Correlation is a broad class statistical relationships between the the relative movement of two variables. Using heatmap generation we find out the correlation between all features. In forecasting crude oil prices, we are using pearson correlation (r) and it measures the linear relationship as well as the degree of relationship with macroeconomic factors.

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2} \sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}} \quad (5.6)$$

Heatmap [Figure 5.13] shows the correlation of all variables.The cutoff of correlation coefficient matrix is 0 because there is no relationship between the pairs. Correlation coefficient pairs which are higher than 0.6 is considered as a highly correlated value. In the heatmap, DIJA and S&P500 are found highly correlated to each other. So, DIJA is dropped. We also considered all negative correlated value.

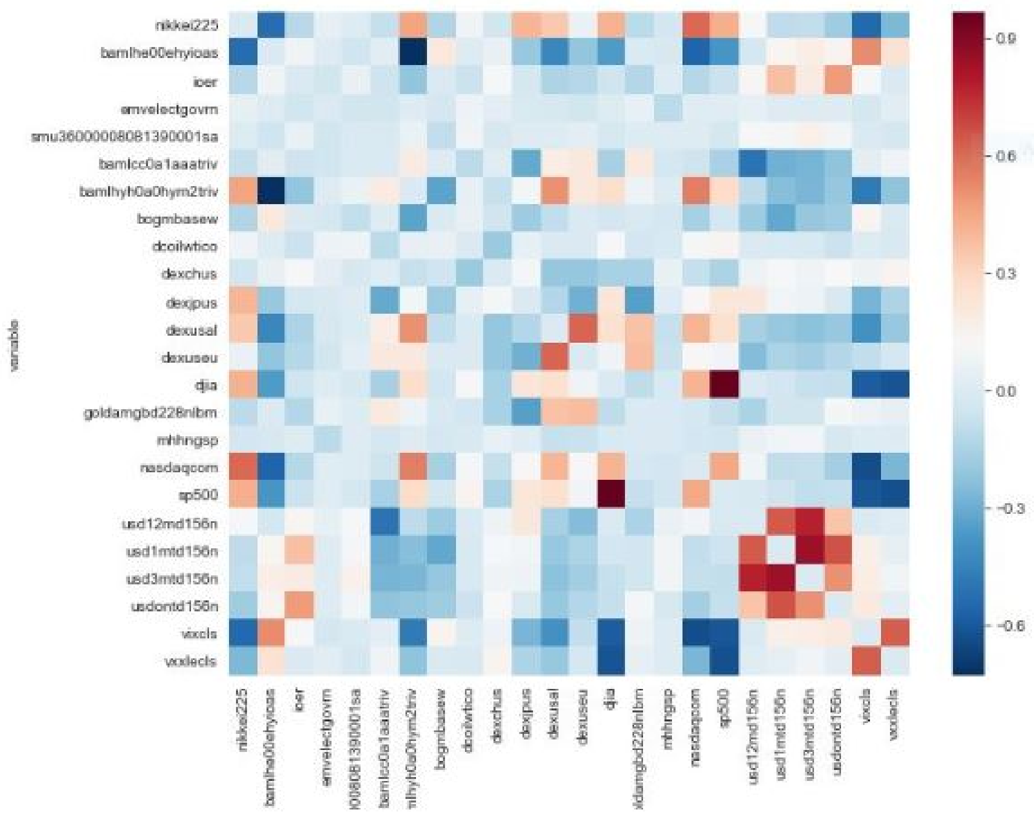


Figure 5.13: Correlation

5.2.6 Selected Features

```
series_list = ['NIKKEI225', 'BAMLHE00EHYIOAS', 'IOER', 'EMVELECTGOVRN', 'SMU36000008081390001SA',
               'SP500', 'NASDAQCOM', 'DJIA', 'BOGMBASEW','DEXJPUS', 'DEXUSEU', 'DEXCHUS', 'DEXUSAL',
              'VIXCLS', 'USDONTD156N', 'USD1MTD156N', 'USD3MTD156N', 'USD12MD156N',
              'BAMLHYH0A0HYM2TRIV', 'BAMLCC0A1AAATRIV',
              'GOLDAMGBD228NLBM',
              'DCOILWTICO',
              'MHHNGSP',
              'VXXLECLS']
```

Figure 5.14: Selected Factors

5.2.7 Preparing Testing and Training data

Our train size was 70% of fetch data and it was tested on 30%. By removing mean and scaling the unit variance standards features were selected. Here we used sklearn.preprocessing.StandardScaler to center the X_cols data around zero. Note that the test data was transformed based on the training data.

Since we were burning through the X_val and y_val data during training we must preserved some data to test our model after it had been trained. For this reason we were splitting the test set in half for final testing on each model

5.2.8 Implementing Feedforward Neural Network

After splitting the dataset for testing and training we feed the data to FNN model for prediction. The loss function for the model is MSE whereas the activation function and optimizer had been chosen relu and rmsprop respectively.

Some terms that is important in FNN are –

Epoch

Epoch is the numerical value which shows the total number of time the training dataset will be traversed by the algorithm. One epoch means the one full traverse through the data.

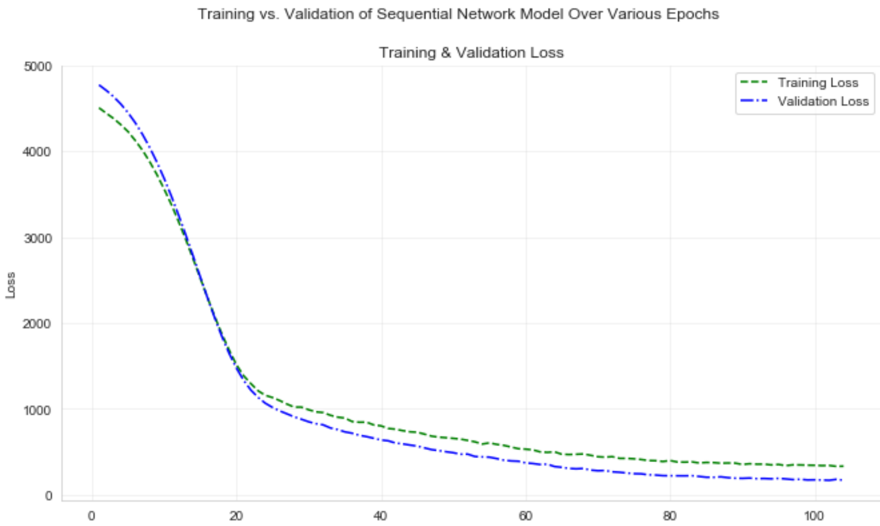
On our model we set an epoch of 100.

Batch

This hyperparameter expound the working sample of the dataset. It defines the number of groups that we divide the samples to work with.

We used batch size of 4 in our model.

The total number of generated model was 96 with RMSE and R squared value on test and train data –




```
Model # 96
Fully Connected Model w/ Dropout & Regularization
 - Regularizer Rate:     0.0005000
 - Dropout Rate:         0.300
 - Number Dense Layers: 2
 - Neurons per Layer:    34
```

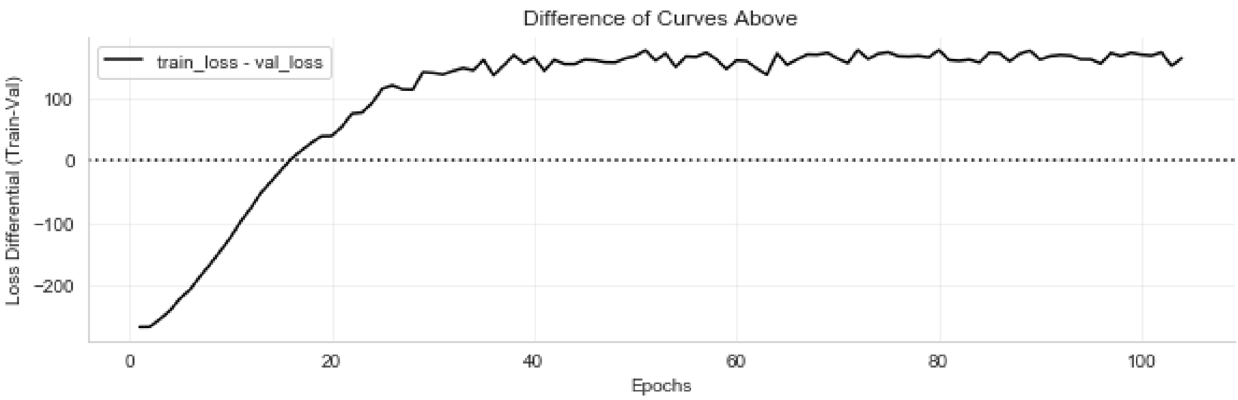


Figure 5.15: Validation and Training Loss for Model 96

Finally the best model has been achieved among the 96 models with greater performance.

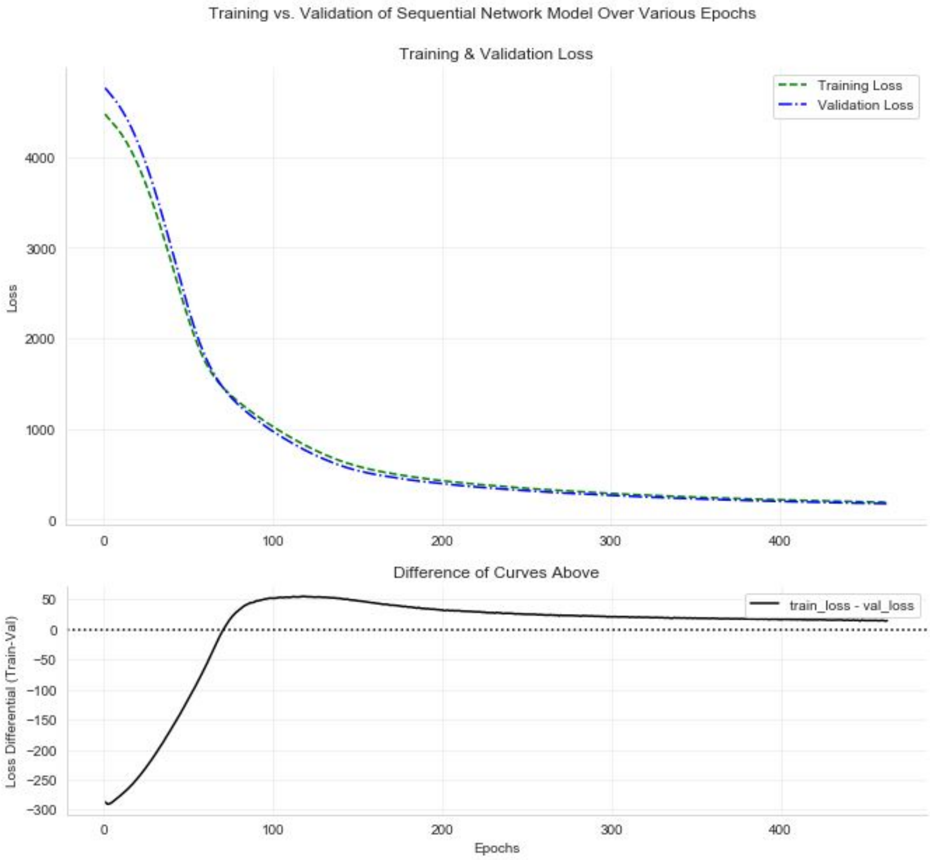


Figure 5.16: Training vs. Validation Loss of Best Model

Chapter 6

Result

6.1 Results from ARIMA

In our approach, prediction using ARIMA was done through considering closing price of daily dataset. With the help of Dickey-Fuller test the non-stationary nature was detected on the given time series. In order to predict the oil market's nature, we determined the p,d,q parameters of ARIMA model.The value of d which makes the time series stationary was determined by using residuals which showed stationary characteristics. The q and p parametric values were determined through the autocorrelation and partial autocorrelation graphs respectively. Lastly, the predicted outcome for the model was –

Figure 6.1: ARIMA prediction

ARIMA model provided good prediction with an RMSE value of 0.8494 and R-squared value of 0.9939.

6.2 Results from FNN

For Feedforward Network we fetched data using Fred API. Geopolitical factors were taken along with macroeconomic and technical factors cause sudden war or any other calamities affect the oil price trend significantly. The macroeconomics factors were taken from the rolling mean and correlation. Cut off value for correlation was 0 thus we could select the factors that had minimum correlation with the market price. To avoid overfitting, we verified the training and testing data through R squared values.

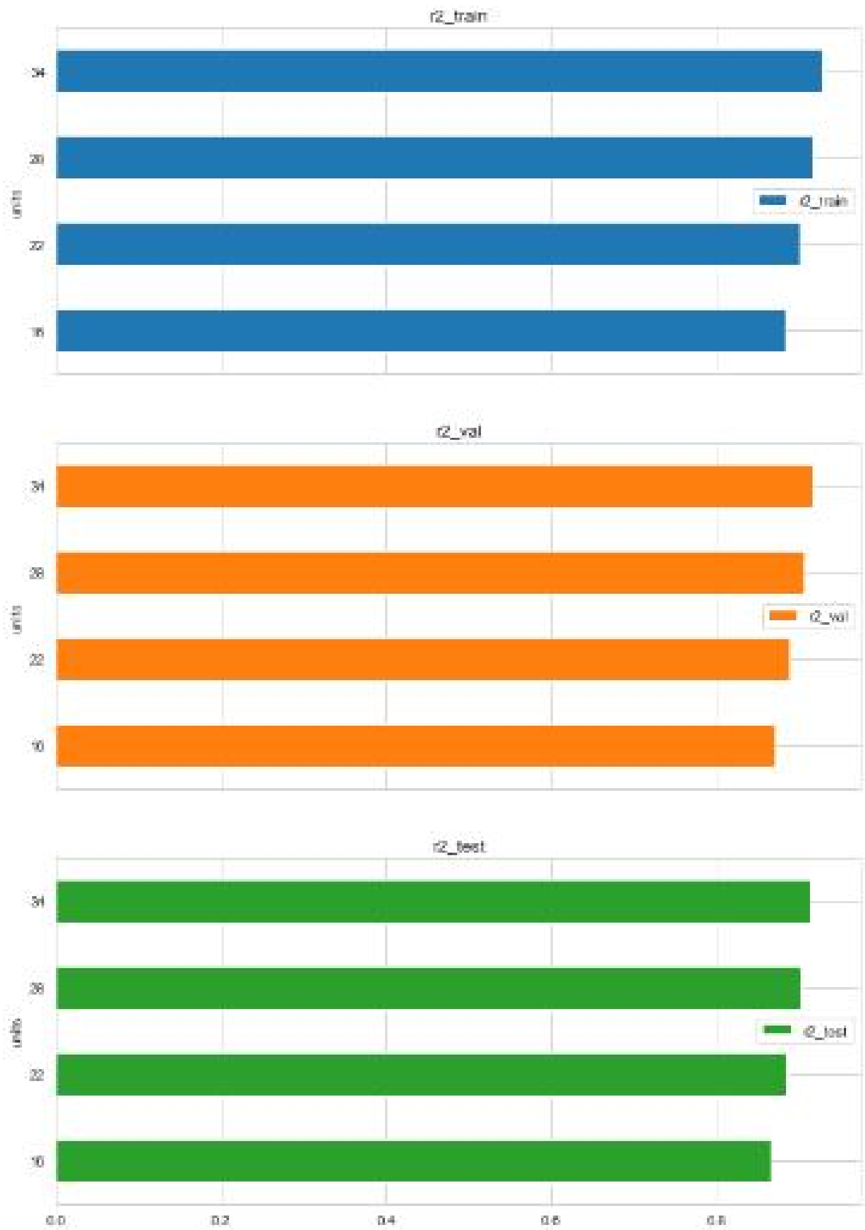


Figure 6.2: R-squared values for training, validation, testing

From the best model the derived R-squared values for training,testing and validation were 0.986209,0.980333, 0.980829 respectively. Through the plotted graphs it can be easily seen that the values are very close to each which indicates that our model does not overfit. Similarly, the values for RMSE are near similar for the train, test and validation with values 3.09585, 3.69149, 3.6648 respectively.

Afterwards, the best model was selected out of 96 generated model through grid search for optimal prediction. The predicted graph for the FNN prediction –

Figure 6.3: Prediction using FNN with actual price

The graph shows the predicted outcome of the model is very close to the actual market price. RMSE and R squared values for FNN are 3.2852 and 0.9847 subsequently.

To show the points of actual and predicted values, we plotted the actual values against the predicted ones and got a linear relationship among them which justified the prediction of FNN was optimal for the generated dataset.

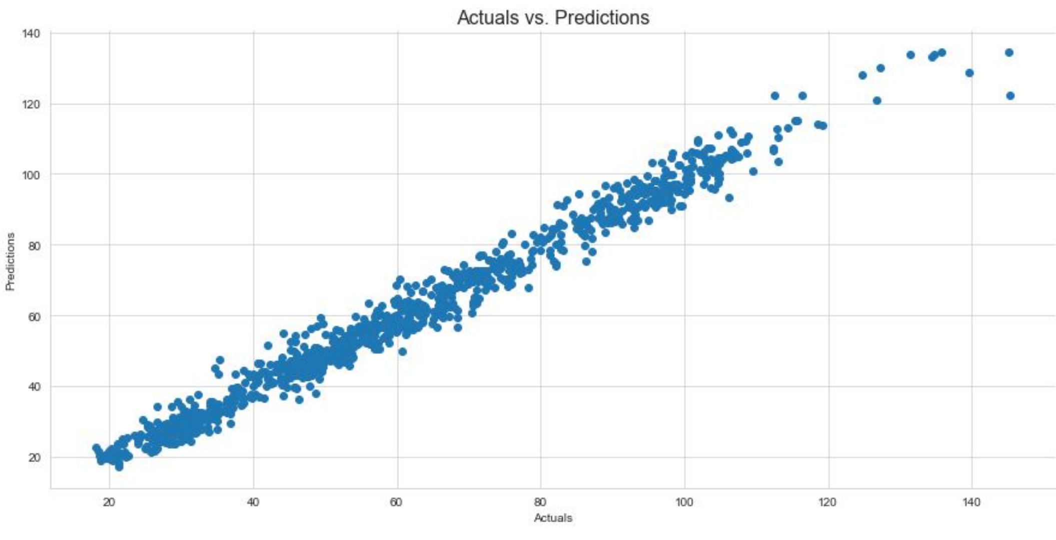


Figure 6.4: Relationship between Actual vs. Prediction

Chapter 7

Comparison

We compared our two main model ARIMA and Feedforward Neural Network with exponential model, SVR and Linear Regression model. The predicted graph for exponential model, SVR and Linear Regression with actual value –

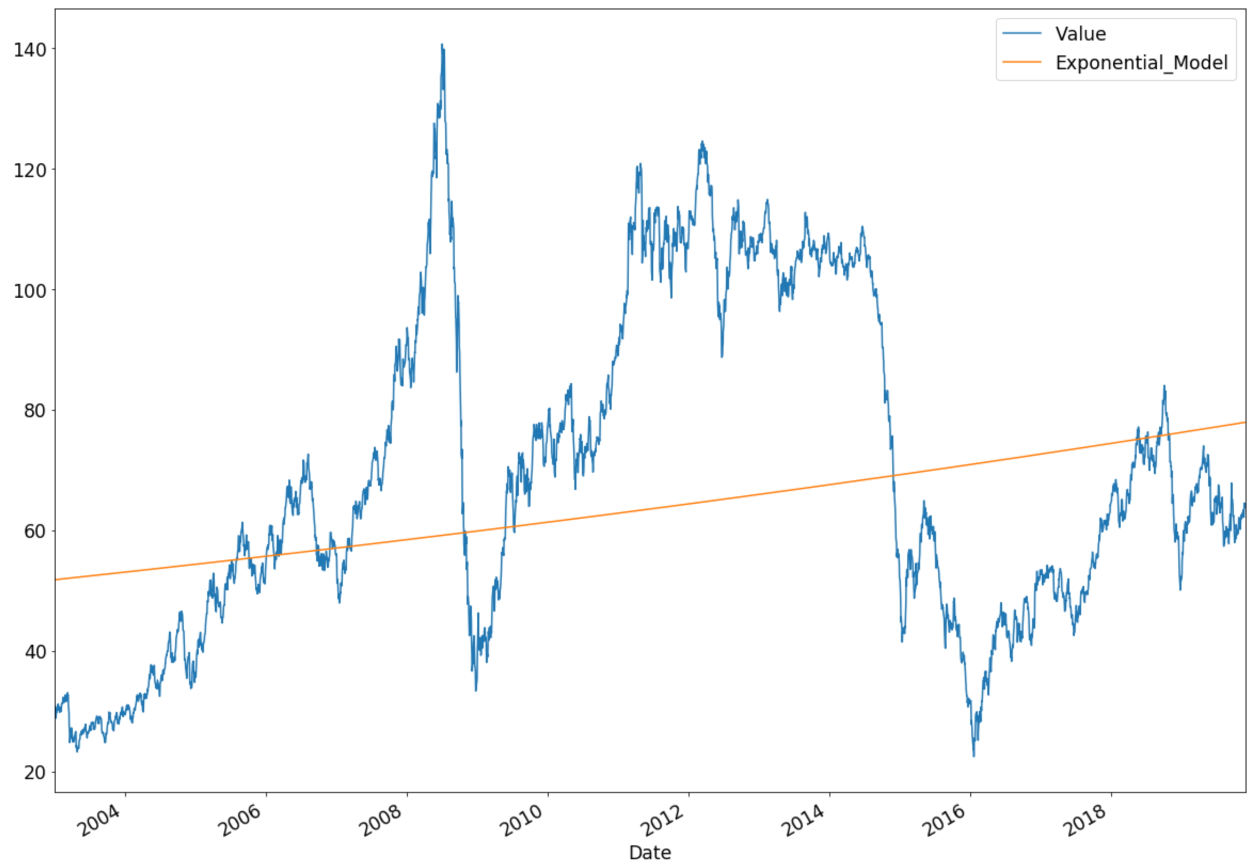


Figure 7.1: Oil price prediction using Exponential model

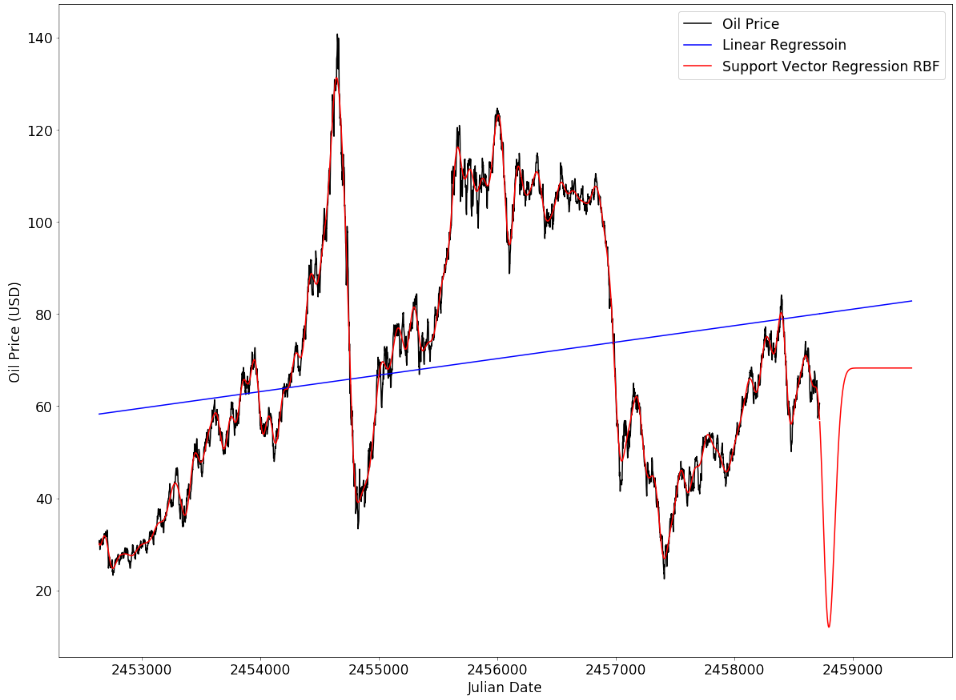


Figure 7.2: Oil price prediction using SVR and Linear Regression

As basic models both exponential and Linear Regression model failed to produce anywhere near to predicted graph hence they gave a linear line for prediction. On the other hand, SVR model predicted reasonably well but could not compete with ARIMA and FNN which can be clearly seen from the predicted output graph.

7.1 Comparison between FNN and ARIMA

Though the ARIMA model had certainly lower value than the FNN in case of RMSE but only lower value of RMSE does not indicate the efficiency of the model. RMSE has the same unit as the dependent variables. Thus, there is no acceptable threshold as it depends on the data representation. In the ARIMA prediction we only used the closing price for the prediction but we used large number of macroeconomic, technical and geopolitical factors to predict value with FNN model. The FNN model definitely worked better with numerous number of factors which can be justified with the R squared value of FNN which was 0.9847.

7.2 Comparison with Previously Implemented Models

In their model Xie at al. (2006) [17] obtained an RMSE value of 2.1921 by implementing Support vector machine. Shabri and Samsudin (2014) [32] implemented their model with Hybridizing Wavelet and Artificial Neural Network (WANN) and their model produce prediction for oil price with an RMSE value for Brent data 0.8151 and for WTI data 0.7449. With their approach using ensemble method Zhao et al. (2017) extract their result with an RMSE value of 4.995 [36]. By implementing LSTM Chen et al. (2017) built their model which gave and MSE of 5.5219 [35]. Varun and Gupta [37] predicted oil price using LSTM and got test and train result for RMSE 2.83 and 5.32 of train and test respectively.

In our FNN model we obtained the RMSE value of 3.2852 and 0.9847 for RMSE and R-squared which showed that the model is good enough for predicting oil market price.

7.3 Limitations

Our model only focused on the data based on United States of America. But to implement a near perfect model we should have considered all the other countries that influence the fluctuations of oil price such as China, Germany, Japan, Russia and so on. We gathered data for 7 countries on 9 factors and tried to use LSTM but could not predict using that dataset because of the data inconsistency occurred for unavailability of data. Moreover, we could not fetch daily data for every factors thus we had to gather the yearly data which resulted in fewer observation which made the prediction failure for FNN.

Chapter 8

Conclusion

8.1 Conclusion

Neural Networks are always good predictors of time series analysis yet the use of Feedforward Neural Network on predicting oil market is not so common. This paper provided an implementation of FNN algorithm in oil price prediction using datasets that were called using API. The macroeconomic, technical and geopolitical factors were selected through rolling mean and autocorrelation. The best model was selected for prediction by using the r squared value of training and testing with justification that the model was not overfitting.

Additionally, ARIMA model with factor of closing price was also used briefly in prediction mechanism. The residuals of the time series were removed to make it stationary from non-stationary thus predicting the outcomes. Finally, the implemented algorithms were compared for performance analysis with known parameters and baseline algorithms. With the outcomes, the research showed the importance of considering the macroeconomic, technical and geopolitical factors in terms of oil price prediction.

8.2 Future work

Recent research showed that the ensemble methods in machine learning worked significantly well in analyzing time series. Our next goal is to use ensemble machine learning along with some mathematical approach to make the model more accurate on different circumstances.

Moreover, we will try to fit more countries and the factors regarding the fluctuations of oil market for better prediction. To do so, our aim is to find data rigorously and find a way to overcome the data inconsistency.

Last and most important aim of ours is to make the algorithm useful for the oil import for Bangladesh thus the general people can get benefited through the efforts that we put.

Bibliography

[1] Dick and H., "The oil price subsidy, deforestation and equity", *Bulletin of Indonesian Economic Studies*, vol. 16, no. 3, pp. 32–60, 1980.

[2] V. Wijnbergen and S., "Oil price shocks, unemployment, investment and the current account: An intertemporal disequilibrium analysis", *The Review of Economic Studies*, vol. 52, no. 4, pp. 627–645, 1985.

[3] Franses and P. H., "Seasonality, non-stationarity and the forecasting of monthly time series. international journal of forecasting", *International Journal of forecasting*, vol. 7, no. 2, pp. 199–208, 1991.

[4] Hagen and R., "How is the international price of a particular crude determined?", *OPEC Review*, vol. 18, 1994.

[5] Huntington and H.G., "Oil price forecasting in the 1980s: What went wrong?", *The Energy Journal*, vol. 15, no. 2, pp. 1–22, 1994.

[6] Svozil, D., Kvasnicka, V., Pospichal, and J., "Introduction to multi-layer feedforward neural networks", *Chemometrics and intelligent laboratory systems*, vol. 39, no. 1, pp. 43–62, 1997.

[7] Gulen and S.G., "Efficiency in the crude oil futures market. journal of energy finance development", *Journal of Energy Finance Development*, vol. 3, pp. 13–21, 1998.

[8] Meyler, A., Kenny, G., Quinn, and T., "Forecasting irish inflation using arima models", 1998.

[9] E. Panas and V. Ninni, "Are oil markets chaotic? a non-linear dynamic analysis", *Energy Economics*, vol. 22, pp. 549–568, 2000.

[10] Kaboudan and M.A., "Compumetric forecasting of crude oil prices", *The Proceedings of IEEE Congress on Evolutionary Computation*, pp. 283–287, 2001.

[11] Lee, B. R., Lee, R. K., and R. A., "Monetary policy, oil price shocks, and the japanese economy", *Japan and the World Economy*, vol. 13, no. 3, pp. 97–112, 2001.

[12] Balke, N. S., Brown, S. P., Yücel, and M. K., "Oil price shocks and the us economy: Where does the asymmetry originate?", *The Energy Journal*, pp. 27–52, 2002.

[13] Bernanke, B., Gertler, M., Watson, and M. W., "Oil shocks and aggregate macroeconomic behavior: The role of monetary policy: A reply", *ournal of Money, Credit, and Banking*, vol. 36, no. 2, pp. 287–291, 2004.

[14] Mirmirani, S., Li, and H.C., “A comparison of var and neural networks with genetic algorithm in forecasting price of oil”, *Advances in Econometrics*, vol. 19, pp. 203–223, 2004.

[15] A. Lanza, Manera, M., Giovannini, and M., “Modeling and forecasting cointegrated relationships among heavy oil and product prices”, *Energy Economics*, vol. 27, pp. 13–21, 2005.

[16] Moshiri, Saeed, Foroutan, and Faezeh, “Forecasting nonlinear crude oil futures prices”, 2006.

[17] Xie, W., Yu, L., S. Xu, Wang, and S., “A new method for crude oil price forecasting based on support vector machines”, *In International Conference on Computational Science, Springer, Berlin, Heidelberg*, pp. 444–451, May 2006.

[18] Shambora, W.E., Rossiter, and R., “Are there exploitable inefficiencies in the futures market for oil?”, *Energy Economics*, vol. 29, pp. 18–27, 2007.

[19] Diao, X., Headey, D., Johnson, and M., “Toward a green revolution in africa: What would it achieve, and what would it require?”, *Agricultural Economics*, vol. 39, pp. 539–550, 2008.

[20] Farzanegan, M. R., Markwardt, and G., “The effects of oil price shocks on the iranian economy”, *Energy Economics*, vol. 31, no. 1, pp. 134–151, 2009.

[21] He, K., Lai, K. K., Yen, and J., “Crude oil price prediction using slantlet denoising based hybrid models”, *2009 International Joint Conference on Computational Sciences and Optimization*, vol. 2, pp. 12–16, Apr. 2009. DOI: 10.1016/j.aasri.2012.06.082.

[22] Jbir, R., Zouari-Ghorbel, and S., “Recent oil price shock and tunisian economy”, *Energy Policy*, vol. 37, no. 3, pp. 1041–1051, 2009.

[23] Rafiq, S., Salim, R., Bloch, and H., “Impact of crude oil price volatility on economic activities: An empirical investigation in the thai economy”, *Resources Policy*, vol. 34, no. 3, pp. 121–132, 2009.

[24] X. Abdullah S. N. & Zeng, “Machine learning approach for crude oil price prediction with artificial neural networks-quantitative (ann-q) model”, *The International Joint Conference on Neural Networks—Proc Int Jt Conf Neural Networks Institute of Electrical and Electronics Engineers*, 2010. eprint: https://doi.org/10.1089/big.2014.0018. [Online]. Available: https://doi.org/10.1109/IJCNN.2010.5596602.

[25] Lizardo, M. R. A., and A. V., “Oil price fluctuations and us dollar exchange rates”, *Energy Economics*, vol. 32, no. 2, pp. 399–408, 2010.

[26] Meyer and P., “Julian day numbers. hermetic systems”, 2010. eprint: http://www.hermetic.ch/cal_stud/jdn.htm. [Online]. Available: http://www.%20hermetic.%20ch/cal_stud/jdn.%20htm.

[27] Ji, Q., Fan, and Y., “How does oil price volatility affect non-energy commodity markets?”, *Applied Energy*, vol. 89, no. 1, pp. 273–280, 2012.

[28] Gabralla, L. A., R. Jammazi, Abraham, and A., "Oil price prediction using ensemble machine learning", *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (IC-CEEE)*, pp. 674–679, Aug. 2013.

[29] Shaari, M. S., Hussain, N. E., A. Rahim, and H., "The effects of oil price changes and exchange rate volatility on unemployment: Evidence from malaysia", *International Journal of Research in Business and Social Science*, vol. 2, no. 4, pp. 7–83, 2013.

[30] Liu and L., "Cross-correlations between crude oil and agricultural commodity markets", *Physica A: Statistical Mechanics and its Applications*, vol. 395, pp. 293–302, 2014.

[31] V. Rathi and S. Aggarwal, "Mcomparing the performance of ann with fnn on mammography mass data set", *IEEE International Advance Computing Conference (IACC), Gurgaon, 2014*, pp. 1307–1314, 2014.

[32] A. Shabri and R. Samsudin, "Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model", *Mathematical Problems in Engineering*, p. 10, 2014.

[33] Wang, Y., Wu, C., Yang, and L., "Oil price shocks and agricultural commodity prices", *Energy Economics*, vol. 44, pp. 22–35, 2014.

[34] Mahdiani, M. R., Khamehchi, and E., "A modified neural network model for predicting the crude oil price", *Intellectual Economics*, vol. 10, no. 2, pp. 71–77, 2016.

[35] K. H. Chen Yanhui and G. K. Tso, "Forecasting crude oil prices: A deep learning based model", *Procedia computer science*, vol. 122, pp. 300–307, 2017.

[36] J. L. Zhao Yang and L. Yu, "A deep learning ensemble approach for crude oil price forecasting", *Energy Economics*, vol. 66, pp. 9–16, 2017.

[37] A. Gupta V. & Pandey, "A. crude oil price prediction using lstm networks",