ISSN 2278-3091

Volume 10, No.6, November - December 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse071062021.pdf

https://doi.org/10.30534/ijatcse/2021/071062021

Smart System for Detecting Anomalies In Crude Oil Prices Using Long Short-Term Memory

P. S. Ezekiel[1], O. E. Taylor[2], M. O. Musa[3]

[1,2]Department of Computer Science, Rivers State University, Port Harcourt, Nigeria.
ezekielpromise27@gmail.com, taylor.onate@ust.edu.ng
[3]Department of Computer Science, University of Port Harcourt, Port Harcourt, Nigeria.
martha.musa@uniport.edu.ng

Received Date : October 10, 2021 Accepted Date : November 09, 2021 Published Date : December 06, 2021

ABSTRACT

Crude oil is leading globally, as it represents roughly about 33% of the total energy consumed globally. It is one of the most significant exchanged resources in the world, oil in one way or the other affects our day to day routines, like transportation, cooking and power, and other numerous petrochemical items going from the things we use to the things we wear. The increment sought after for petroleum derivatives is on a persistent ascent, making it vital for the oil and gas industry to think of new methodologies for further developing activity. This paper presents a smart system for detecting anomalies in crude oil prices. The experimental process of the proposed system is of two phases. The first phase has to do with the pre-processing stage, and the training stage while the second phase of the experiment has to do with the building/training of the Long Short-Term Memory algorithm. The experimental result shows that LSTM model had an accuracy result of 98%. The result further shows that our proposed model is under fitting since the training loss is lesser than the validation loss. The proposed model was saved and was used in detecting anomalies of the crude oil prices ranging from 1990 to 2020.

Key words : Anomalies, Crude Oil Price, Deep Learning, Long Short-Term Memory

1. INTRODUCTION

Crude oil is leading globally, as it represents roughly about 33% of the total energy consumed globally. It is one of the most significant exchanged resources in the world, oil in one way or the other affects our day to day routines, like transportation, cooking and power, and other numerous petrochemical items going from the things we use to the things we wear. The increment sought after for petroleum derivatives is on a persistent ascent, making it vital for the oil and gas industry to think of new methodologies for further developing activity. The oil and gas is among the biggest, mind boggling and pivotal enterprises on the planet, with series of upstream, halfway and downstream exercises occurring, including creation, refining, transportation and utilization. The business is additionally portrayed by working and legally binding associations with states, joint endeavors and different partners. It is exceptionally unique in nature, making it hard for the customary unified way to deal with work, particularly with the ascent in worldwide interest. Endeavors to recognize imaginative ways to deal with process and oversee data for the motivations behind improving functional strength, decreasing working expenses and supporting benefits require headways in innovations inside the business [1].

As the most crucial non-environmentally power on planet earth, raw petroleum assumes a critical and indispensable part in monetary society. Raw petroleum is a significant material for some, synthetic modern items including manures, solvents, plastics and pesticides. The costs of worldwide raw petroleum rest with the world's significant oil-creating regions. Numerous business analysts recommend that high price of oil contrarily affect worldwide monetary development, while others believe that high price of oil are brought about by financial development, as a rule, the connection between prices of oil and worldwide monetary is entirely unsound [2]. The elements of oil costs influence the degree of expenses in all areas of the world. The economy of many nations depends on oil creation and exchange oil and oil items, accordingly, predicting the prices of oil is a significant task. It is likewise significant that a few areas of the economy are straightforwardly subject to oil costs. Prices of oil influence political and monetary cycles that decide the worth of oil organizations' stocks, the pace of expansion in the nations that import oil, and the pace of financial development. Note the effect of prices of oil on the valuing of elective energy sources [3].

The world's current circumstance is impacted by the fluctuation of oil and gas prices. With the drop in prices of oil, the fuel bills are brought down. Subsequently, buyers are probably going to utilize more oil and consequently increment the fossil fuel byproduct. Likewise, there is less motivating force to foster sustainable and clean energy assets. Then again, supported low oil costs could prompt a widely drop in exploring the activities of oil and gas. Fluctuation in the prices of oil also perform a significant part in a widely scale economy [4]. The fall in the prices in oil would bring about a humble lift to worldwide monetary action, albeit the proprietors of oil areas endure pay misfortunes. Ongoing exploration from the World Bank shows that for each 30% decrease of oil costs, the worldwide GDP (Gross Domestic Product) would be expanded by 0.5%. Simultaneously, the drop of oil costs would diminish the average cost for basic items, and thus the expansion rate would fall. Beside Supply and Demand Fundamentals, various factors strike the oil costs at various frequencies. In energy market, creation of different products including flammable gas, coal and sustainable power, may have replacement impact which prompts the unpredictability of oil cost by implication. Different factors like monetary business sectors, financial development, innovation improvement and sporadic occasions, additionally impact the prices of oil in various ways. Complex connections are worked between these elements and prices of oil, thus, driving a solid variance in crude oil market. To determine the prices of oil has consistently been an extreme approach. Nonetheless, looking for encouraging forecasting techniques for oil value series is not really obsolete since unrefined petroleum is the fundamental wellspring of energy on the planet and overwhelm the monetary exercises. Precise figure of oil value directs the dynamic of numerous areas like business associations and state run administrations [5].

2. RELATED WORKS

The research [1] summarized the modern endeavors of applying multi-agents system and machine learning framework in predicting the prices of two oil and gas industries, recognizing potential purposes behind their low and slow take-up and recommend ways of guaranteeing a more prominent technique to ensure a higher acquisition in the oil and gas industry. Their summary revealed that machine learning frameworks can be a significant instrument for mining examples and data from the information and creating forecasts to help in making decision, while multi specialist framework can possibly assist with dealing with the dynamic, circulated and questionable connections in the business.

The paper [3] introduced a modified linear regression based model for foreseeing the prices of oil and gas. They tested their model on a dataset that contains Brent oil prices, and their model showed a promising accuracy when prices of Brent oil falls. Though their proposed model did not consider the impact of outside factors, such as market crisis. The results of their model improved the value of relative mean square error to 994.38, and mean absolute error to 927.76.

The paper [6] proposed a modified hybrid modeling framework in making analysis of the prices in oil and gas. Their proposed framework was gotten from the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model and LSTM (long-transient memory). They applied their proposed framework to the prices of crude oil in two phases. the first phase was where system advances are administered in both the restrictive mean and the contingent difference, while the second phase was a correlation of their proposed technique with the GARCH and LSTM strategies predicting the prices of oil. Their proposed framework achieved a higher accuracy over the others.

In their research, [7] proposed a deep learning framework to catch the obscure complex nonlinear attributes of the value of crude oil. Their proposed hybrid model was trained using a deep learning algorithm in forecasting the prices in crude oil using. The model. Their experimental result shows that the performance of the proposed model is evaluated using the price data.

In the research [8], a framework in forecasting the prices of crude oil was presented. The proposed framework was trained using Long Short-Term Memory. They also applied an empirical ensemble mode in decomposing time series into a different natural mode capacity, and these inborn mode capacities were used in training the model. They tested the forecast impact of West Texas Intermediate and Brent raw petroleum on their model. This was archived by assessing the foreseeing capacity of their proposed model, demonstrating the corresponding superiority of their model.

The paper [9] proposed a novel methodology in predicting the prices of crude oil using stream learning. The stream learning was utilized in catching changes in the prices of crude oil since the model is persistently refreshed at whatever point new oil value information are accessible, with tiny steady overhead. Their experimental results show that the proposed stream learning model accomplishes the most elevated precision as far as both mean squared expectation blunder and directional accuracy proportion over a predictive time horizon.

In the research [10], a framework in foreseeing raw petroleum costs utilizing XGBoost was proposed. The proposed XGBoost framework was trained on a dataset that involves crude oil value, gold, silver and flammable gas. Their experimental result shows an improved performance of their model on a validation test.

The research [11] proposed a deep learning framework name stacked denoising autoencoder and bootstrap aggregation to foresee raw petroleum cost. The stacked denoising autoencoder was utilized to display the non-straight complex connections of oil cost with its variables, while the bootstrap was utilized in preparing various datasets for their based model. The result of their experimental result shows their proposed model was outstanding on three test data.

3. DESIGN METHODLOGY

These section describe the processes involved in training the Deep Learning model in detecting anomalies in crude oil prices. The architecture of the proposed system can be seen below in figure 1.

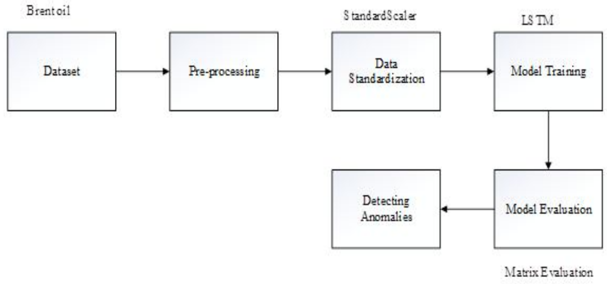
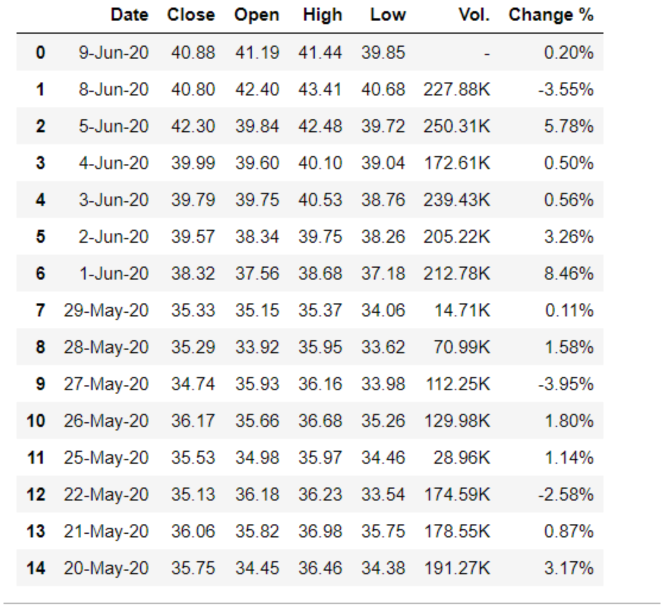


Figure 1: Architecture of the proposed system

Dataset: The Brent oil benchmark dataset will be used in training our Long Short-Term Memory Algorithm. The dataset comprises of 6 columns stating from Date columns and the closing price column. The dataset was download from kaggle.com. Dataset sample can be seen in figure 2



	Date	Close	Open	High	Low	Vol.	Change %
0	9-Jun-20	40.88	41.19	41.44	39.85	-	0.20%
1	8-Jun-20	40.80	42.40	43.41	40.68	227.88K	-3.55%
2	5-Jun-20	42.30	39.84	42.48	39.72	250.31K	5.78%
3	4-Jun-20	39.99	39.60	40.10	39.04	172.61K	0.50%
4	3-Jun-20	39.79	39.75	40.53	38.76	239.43K	0.56%
5	2-Jun-20	39.57	38.34	39.75	38.26	205.22K	3.26%
6	1-Jun-20	38.32	37.56	38.68	37.18	212.78K	8.46%
7	29-May-20	35.33	35.15	35.37	34.06	14.71K	0.11%
8	28-May-20	35.29	33.92	35.95	33.62	70.99K	1.58%
9	27-May-20	34.74	35.93	36.16	33.98	112.25K	-3.95%
10	26-May-20	36.17	35.66	36.68	35.26	129.98K	1.80%
11	25-May-20	35.53	34.98	35.97	34.46	28.96K	1.14%
12	22-May-20	35.13	36.18	36.23	33.54	174.59K	-2.58%
13	21-May-20	36.06	35.82	36.98	35.75	178.55K	0.87%
14	20-May-20	35.75	34.45	36.46	34.38	191.27K	3.17%

Figure 2: Brent oil dataset

Pre-processing: The pre-processing has to do with the data cleaning. By data, we mean the checking and removal of null value, converting the column to normal date format.

Feature extraction: Feature extraction was used in selecting the most appropriate features that is suitable in training our LSTM model. Out of the six columns, we selected just two columns. The selected columns are the Date column and the Class column. The Date and Class columns contains the prices of crude oil price starting from 1998 to 2020.

Data Standardization: Data standardization was used in transforming and bringing the dataset to a standard form. For data standardization, we will be making use of StandardScaler() in transforming and bringing our training data to a standard format.

Model Training: For our model training, we will make of Long Short-term Memory algorithm in training our proposed model. The LSTM is a Recurrent Neural Network algorithm. The LSTM model will be trained utilizing Tensorflow Framework with Keras application. Keras Sequential API which implies we develop the organization each layer in turn. The layers are as per the following:

1. An Embedding, which maps each info word to a 100-dimensional vector. The implanting can utilize pre-prepared weights (more in a second) which we supply as parameters.
2. A Masking layer to veil any words that don't have a pre-prepared implanting which will be represented as 0s. This layer ought not be utilized when preparing the embeddings.
3. The heart of the organization: a layer of LSTM cells with dropout to forestall overfitting. Since we are just utilizing one LSTM layer, it doesn't return the successions.
4. A completely associated Dense layer with relu enactment. This adds extra illustrative ability to the organization.
5. A Dropout layer to forestall overfitting to the preparation information.
6. A Dense completely associated yield layer. This creates a likelihood for each word in the vocab utilizing softmax enactment.

The architecture of LSTM can be seen in figure 3

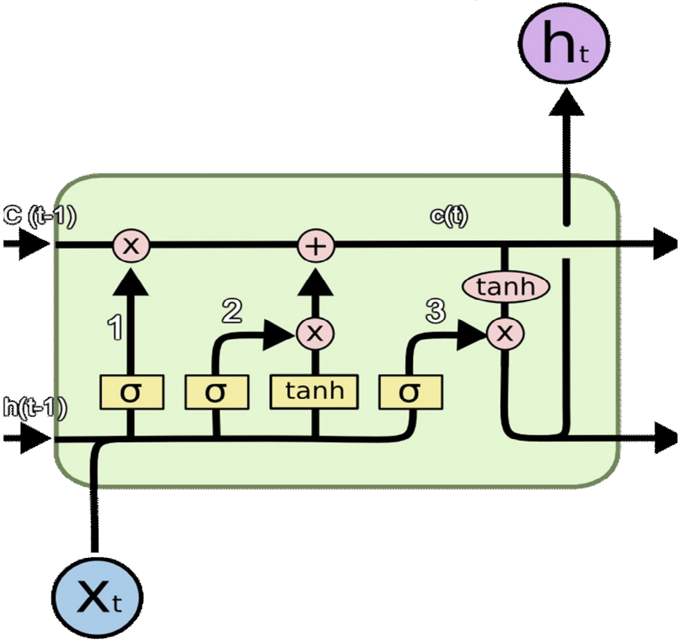


Figure 3: Architecture of Long-Short Term Memory Algorithm

Model Evaluation: The evaluation matrix is used in evaluating the performance of the model in terms of mean absolute error, validation test, accuracy and loss. Mean absolute error (MAE) is a proportion of errors between matched perceptions communicating a similar peculiarity. The mean outright mistake is a typical proportion of estimate error in time series analysis, here and there utilized in disarray with the more standard meaning of mean absolute deviation..

Detecting Anomalies: The trained model will be used in detecting anomalies in crude oil prices across 1990 to 2020.

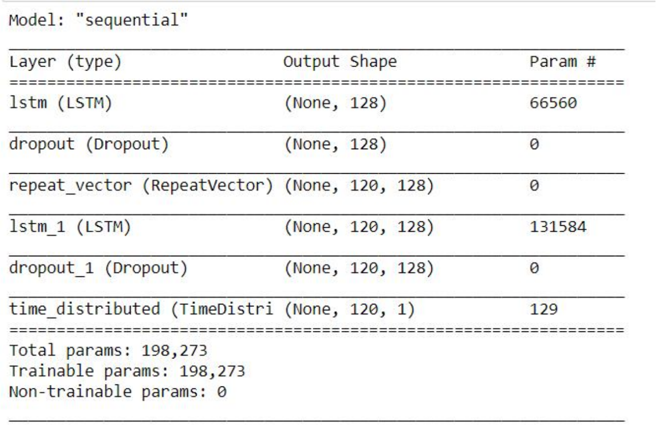
4. Result and Discussion

This system presents a smart system for detecting anomalies in crude oil prices. The experimental process of the proposed system is of two phase. The first phase has to do with the pre-processing stage, and the training stage. In the first stage, the system starts by acquiring a crude oil dataset, performing some pre-processing and building a model using Deep Learning Algorithm. The dataset used here is Brent oil price dataset. The dataset comprises of seven columns starting from Date, Close, Open, High, Low, Volume, and Change. The dataset contains crude oil prices starting from 1998-2020. The dataset was pre-processing by checking for null values and as well dropping rows that contains missing or Nan values. This was achieved using pandas library in calling the function dataset.insull() for checking for missing values and dataset.dropna() in dropping rows that contains missing values. After pre-processing, we applied feature extraction technique in selecting just few columns that are important in training our proposed model in detecting anomalies in crude oil prices. The selected features are the date column and the Close column. Sample of the selected features can be seen in table 1. After the process of feature extraction, we used StandardScaler() in scaling and bringing our training data to a standard and balanced format. The second phase of the experiment has to do with the building/training of the model. The model was trained using Long Short-Term Memory algorithm (LSTM). The number of layers used in training the LSTM model is two. The e input layer, which is the first layer, is made of 128 neurons, and the second layer is the output layer (one output). The parameters used in training the LSTM model are dropout = 0.2, optimizer: Adam with learning_rate = 0.01, validation_split = 0.2, batch_size = 32, epochs = 100, the model summary before training is shown in figure 4. Figure 5 shows the training process of carried out by the LSTM model for the first ten steps. The training process comprises of the loss values gotten during training and validation data. Figure 6 shows the training validation loss of the model on 100 training steps. After training, the model was saved and evaluated. The evaluation performance of the model shows that the model had an accuracy of about 98%. Figure 7 shows a distribution plot of the training loss against density. Figure 8 and 9 shows a graphical representation of the crude oil prices and anomalies for 1998-2020. Table 1 shows the anomalies detected in the crude oil prices for the first ten rows.

Table 1: Training Data

S/N	Date	Close
1	1998-07-15	14.8
2	1998-07-14	14.06
3	1998-07-13	14.17
4	1998-07-12	14.10
5	1998-07-11	14.50
6	1998-07-08	15.05
7	1998-07-07	15.45
8	1998-07-06	14.68
9	1998-07-05	14.30
10	1998-07-04	14.35
11	1998-07-01	14.60
12	1998-06-30	14.85
13	1998-06-29	15.47
14	1998-06-28	15.27
15	1998-06-27	15.10
16	2020-06-09	40.88
17	2020-06-08	40.80
18	2020-06-05	42.30
19	2020-06-03	39.79
20	2020-05-29	35.33

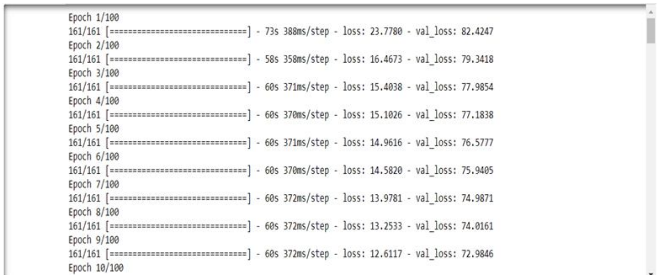
Table 1 shows the extracted features gotten from the original dataset (which comprises of six columns). This are sample data of Brent crude oil starting from 1998 to 2020. The data column contains the number of years starting from 1998 to 2020, whereas the close column contains prices of the crude oil of each month of the year.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 128)               66560
_________________________________________________________________
dropout (Dropout)            (None, 128)               0
_________________________________________________________________
repeat_vector (RepeatVector) (None, 120, 128)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 120, 128)          131584
_________________________________________________________________
dropout_1 (Dropout)          (None, 120, 128)          0
_________________________________________________________________
time_distributed (TimeDistri (None, 120, 1)            129
=================================================================
Total params: 198,273
Trainable params: 198,273
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: Model Summary

This shows the summary of the Long Short-Term Memory Model. The model summary shows the parameters that will be used in training the model. It shows a total number of 120 input neuron and one output layer.



```
Epoch 1/100
161/161 [==============================] - 73s 388ms/step - loss: 23.7780 - val_loss: 82.4247
Epoch 2/100
161/161 [==============================] - 58s 358ms/step - loss: 16.4673 - val_loss: 79.3418
Epoch 3/100
161/161 [==============================] - 60s 371ms/step - loss: 15.4038 - val_loss: 77.9854
Epoch 4/100
161/161 [==============================] - 60s 370ms/step - loss: 15.1026 - val_loss: 77.1838
Epoch 5/100
161/161 [==============================] - 60s 371ms/step - loss: 14.9616 - val_loss: 76.5777
Epoch 6/100
161/161 [==============================] - 60s 370ms/step - loss: 14.5820 - val_loss: 75.9405
Epoch 7/100
161/161 [==============================] - 60s 372ms/step - loss: 13.9781 - val_loss: 74.9871
Epoch 8/100
161/161 [==============================] - 60s 372ms/step - loss: 13.2533 - val_loss: 74.0161
Epoch 9/100
161/161 [==============================] - 60s 372ms/step - loss: 12.6117 - val_loss: 72.9846
Epoch 10/100
```

Figure 5: Sample of training steps

This shows the training process for the first 10 process. The training process shows the time taken to complete one training step, the loss value gotten while training the model at that

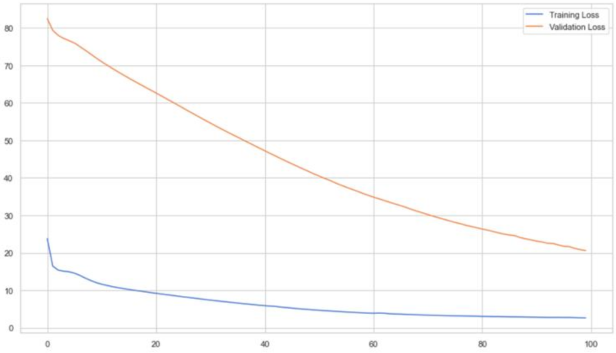


Figure 6: Training loss and Validation.

The line representation of the training loss and validation loss described the fittings of the model. The blue line which indicates the training loss decreases from 25 at the first step and 2 at the last step being 100. This shows that the model is in good performance. The red line, which indicates the validation test, shows that the loss value started from 83 and ended at 20. Therefore since the loss value of the training loss is lesser than the loss value of the validation model, therefore, this shows that the training model is under fitting. Which means the model is in good shape.

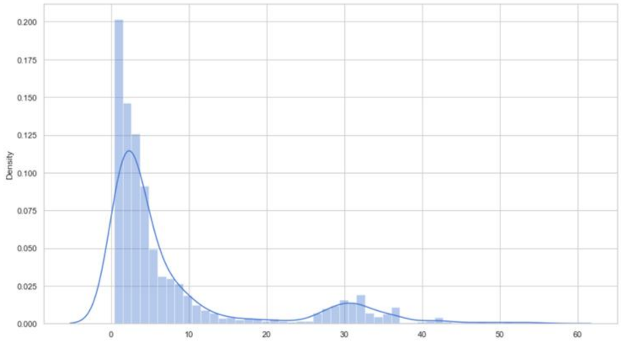


Figure 7: Distribution plot of Mean Absolute Error

The mean absolute error shows the average over the test of the total changes between actual forecast perception where all singular changes have distinctive weight.

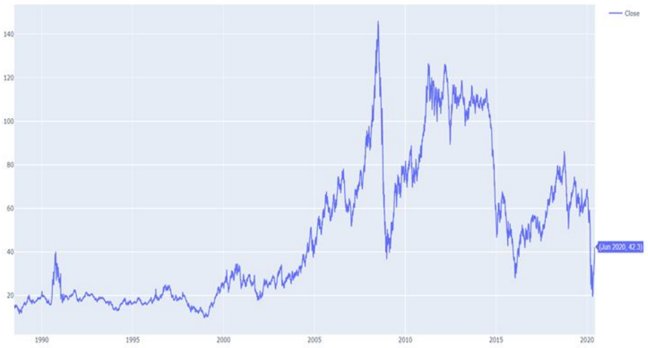


Figure 8: graphical representation of crude oil price

Figure 8 shows the graphical representation of oil prices against time. This shows the fluctuation in oil prices starting from 1990 to 2020

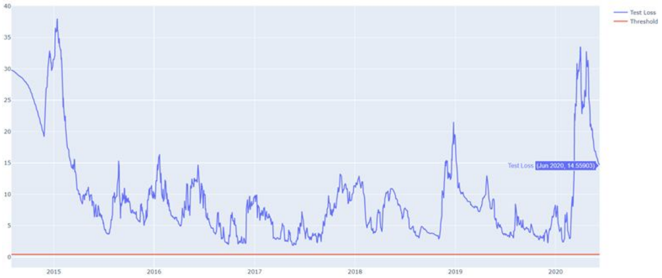
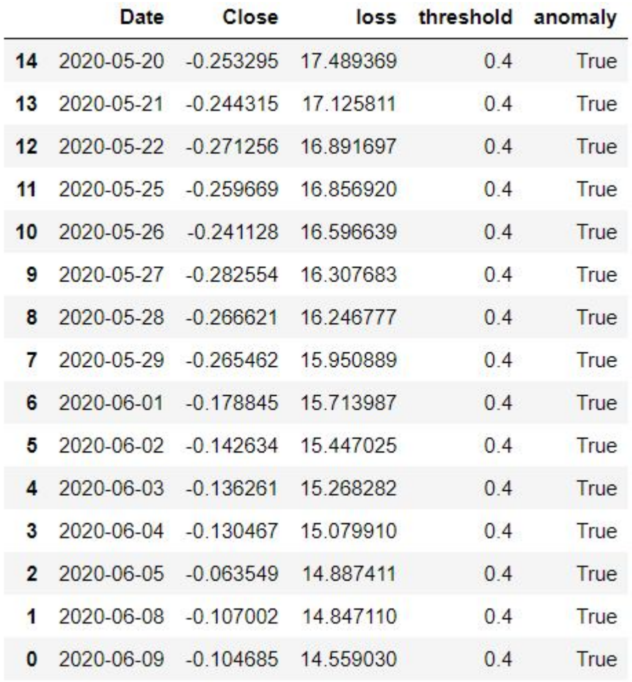


Figure 9: graphical representation crude oil price, test loss, threshold and time

In figure 9, the test loss represents the anomalies gotten from the crude oil price starting from 1990 to 2020.



	Date	Close	loss	threshold	anomaly
14	2020-05-20	-0.253295	17.489369	0.4	True
13	2020-05-21	-0.244315	17.125811	0.4	True
12	2020-05-22	-0.271256	16.891697	0.4	True
11	2020-05-25	-0.259669	16.856920	0.4	True
10	2020-05-26	-0.241128	16.596639	0.4	True
9	2020-05-27	-0.282554	16.307683	0.4	True
8	2020-05-28	-0.266621	16.246777	0.4	True
7	2020-05-29	-0.265462	15.950889	0.4	True
6	2020-06-01	-0.178845	15.713987	0.4	True
5	2020-06-02	-0.142634	15.447025	0.4	True
4	2020-06-03	-0.136261	15.268282	0.4	True
3	2020-06-04	-0.130467	15.079910	0.4	True
2	2020-06-05	-0.063549	14.887411	0.4	True
1	2020-06-08	-0.107002	14.847110	0.4	True
0	2020-06-09	-0.104685	14.559030	0.4	True

Figure 10: Sample Table of Anomaly, loss, closing price and Date.

Figure 10 shows the distortions in returns that contradict the efficient market hypothesis. The pricing anomalies shows how the crude oil is priced differently than how the model predicts it will be priced.

5. CONCLUSION AND FUTURE WORK

This system presents a smart system for detecting anomalies in crude oil prices. The experimental process of the proposed system is of two phases. The first phase has to do with the pre-processing stage, and the training stage while the second phase of the experiment has to do with the building/training of the Long Short-Term Memory algorithm. The experimental result shows that LSTM model had an accuracy result of 98%. The result further shows that our proposed model is under fitting since the training loss is lesser than the validation loss. The proposed model was saved and was used in detecting anomalies of the crude oil prices ranging from 1990 to 2020. This work can further be extended by training different deep learning algorithm and comparing each of them, in other to have a more better accuracy result. The world's current circumstance is impacted by the fluctuation of oil and gas prices. With the drop in prices of oil, the fuel bills are brought down. Subsequently, buyers are probably going to utilize more oil and consequently increment the fossil fuel byproduct. Likewise, there is less motivating force to foster sustainable and clean energy assets.

REFERENCES

[1]. M. K. Hanga, Y. Kovalchuk “Machine learning and multi-agent systems in oil and gas industry applications: A survey”, Computer Science Review, 34(2019) 1-19 2019.

[2]. H. Chiroma, S. Abdulkareem, T. Herawan, “Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction”, Applied Energy, 142(1), 266-273, 2015.

[3]. A. Jaehyung , A. Mikhaylov, N. Moiseev, “Oil Price Predictors: Machine Learning Approach”, International Journal of Energy Economics and Policy, 9(5), 1-6, 2019.

[4]. A.M. Husain, R. Arezki, P. Breuer, V. Haksar, M. Thomas, “Global implications of lower oil prices”, International Monetary Funds Staff Discussion Notes 1-41, 2015.

[5]. C. Baumeister, L. Kilian, “Forty Years of Oil Price Fluctuations: Why the Price of Oil May Still Surprise Us”. Journal of Economic Perspectives, 30(1), 139-160, 2016.

[6] M. Bildirici, N. G. Bayazit, Y. Ucan “Analyzing Crude Oil Prices under the Impact of COVID-19 by Using LSTARGARCHLSTM”, Energies 13(11), 1-18, 2020.

[7]. Y. Chen, H. Kaijian, K.F. Geoffrey “Forecasting Crude Oil Price: A Deep Learning Based Model”, Information Technology and Quantitative Management, 112(2017), 300-307, 2017.

[8]. Z. Cen, J. Wang, “Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer”, Energy 169(2019), 160-171, 2019.

[9]. S. Gao, Y. Lei, “A new approach for crude oil price prediction based on stream learning”, Geoscience Frontiers, 30(2016), 1-5.

[10]. M. Gumus and M. Kiran, “Crude Oil Price Forecasting Using XGBoost”, 2nd International Conference on Computer Science and Engineering, 2017.

[11]. Y. Zhao, J. Li, Y. Lean, “A deep learning ensemble approach for crude oil price forecasting”, Energy Economics 2017, http://dx.doi.org/10.1016/j.eneco.2017.05.023.