Machine Learning Algorithms for Oil Price Prediction

J Shiva Keerthan, Y Nagasai, Subhani Shaik

ABSTRACT Crude oil is world's most leading fuel. Some machine learning models fits the dataset efficiently depending upon the type of datapoints provided. The main aim of this project is to find the different models that efficiently fit the datapoints and predict the price of fuel with the help of machine learning model[5]. This project aims to compare the different supervised learning models and bring a conclusion based on the efficiency. We have used 5 supervised learning models SVR(linear,RBF,polynomial),RandomForestRegressio-n,Linear Regression, to know which gives best in terms of accuracy and performance we have tried these algorithms which are mostly adaptive to many environments. Now-a-days the oil price has been increasing in leaps and bounds due to certain reason like inflation throughout the world. This has become a major problem in India where prices of LPG (Liquified Petroleum Gas), Petroleum, Diesel have been increasing. Hence these are derived or extracted from crude oil; India gets its source of crude oil from neighbouring countries like Dubai and Saudi-Arabia. To predict the values of the petroleum and Diesel in the mere future, we have decided to use the Machine Learning algorithms and after choosing set of algorithms we have chosen the Linear Regression algorithm, which have given the most accurate results.

Keywords: Prediction, Oil Prices, Machine Learning Models

I. INTRODUCTION

As we are experiencing an unstable increase in prices of oil prices where the oil prices are dependent on the crude oil prices of Dubai and Saudi-Arabia. The transportation will be affected for change in prices. The crude oil price of India has also been taken into consideration for the accurate prediction of prices. I have used various algorithms for predicting the diesel price in India. The algorithms which we have used are Random Forest, Support Vector Regression (RBF model, polynomial model, and linear model), and Linear Regression. The Prediction of crude oil rates based on the previous datasets on the data and prices as the feature list are inputs and target list are predicted values. The implementation was on the logistic regression model which is feasible to some extend for the prediction of the crude oil prices. The implementation is on predicting the crude oil prices for the days using linear regression Python machine learning Algorithm and plotting the graph based on prediction.This paper is investigated as follows. In next section, we discussed about the proposed work, Section 3 discusses dataset

II. PROPOSED METHOD

We have chosen Linear Regression which is the best fit among Random Forest, Support Vector Regression (rbf model, polynomial model, and linear model), and Linear Regression. The predictions are most approximate with Linear Regression Algorithm. The algorithm automatically uses the kernel function that is most appropriate to the data. SVR uses the linear kernel when there are many attributes (approx. 100) in the training data, otherwise it uses the Gaussian Kernel. In the proposed system we have taken the datasets which has the Crude oil price and diesel price. Based on the dataset we have made feature list and target list where the target list is price value of diesel and feature list is the Crude oil price. After the analysis of data is done we have fitted both feature list and target list using python Machine Learning Linear Algorithm and predicted the values for feature list from the dataset values. Finally we have plotted a graph based on the results.

A (a). Attribute selection & collecting data

After making an analysis of what data to be collected according to the prediction of oil price we have finalised the first attribute to be crude oil price in the place from where we import the crude oil (Dubai). Where the transportation also makes a difference in oil price we have also taken the crude oil price of India relating to crude oil price of Dubai. So, we have finalised the attributes which are crude oil price in Dubai and crude oil price of India. we have also taken diesel price of Dubai into consideration which may help in predicting the diesel price of India accurately. We collected the data from the website (www.indexwiz.com). After the attribute selection data cleaning is done according to the dataset which we have collected like removing the Null values. The mandatory thing which we need to after collecting the attributes is to know the correlation among the attribute

A(b). Finding Correlation between the attributes

$$\boldsymbol{r_{xy}} = \frac{\boldsymbol{n\sum x_i y_i - \sum x_i \sum y_i}}{\sqrt{\boldsymbol{n\sum x_i^2 - (\sum x_i)^2}}\sqrt{\boldsymbol{n\sum y_i^2 - (\sum y_i)^2}}}$$

A (c). Finding bias and variance for the attribute values

The bias is produced in the machine learning algorithms due to erroneous assumptions made on the dataset while generating patterns.

The bias can be calculated by:

$$\text{Bias } [\hat{f}(\text{x})]= \text{E } [\hat{f}(\text{x})] - f(\text{x})$$

The variance is error that can cause the algorithm to model the noise present in the dataset.

Hence it can be calculated by:

$$\text{Var } [\hat{f}(\text{x})]= \text{ E}[\ \hat{f}(x)^2 \text{ - E } [\hat{f}(\text{x})]^2$$

Revised Manuscript Received on June 05, 2019

J Shiva Keerthan, Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Y Nagasai, Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Dr.Subhani Shaik, Associate Professor, Department of IT, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Where $\hat{f}(x)$ is the target value predicted by the algorithm and the f(x) is the the original target value.

B. Selecting the accurate model and fitting the model

After a precise dataset is prepared, the dataset is divided into testing set and training set accordingly where the division is done accordingly as per the universal rule (80-20 rule). This rule says that the dataset is divided into 80% of the training set and 20% for the test set. after the division of the dataset into test set and training set we have randomly selected the 3 models which perform well on unseen data, they are SVR, Random Forest, Linear Regression. the fitting is done for the each model and results are to be taken from the model which builds up accurately predicting the precise values of oil.

C. Predicting and plotting the results

After fitting the model the accuracies are to be known for training and testing and the we can proceed with predicting the test data only after the selected model training accuracy is high among the selected model. Hence we have gone with linear regression which has high training accuracy. After selecting the model with high training accuracy we can proceed with predicting the test values. After testing set is predicted the results are to be plotted.

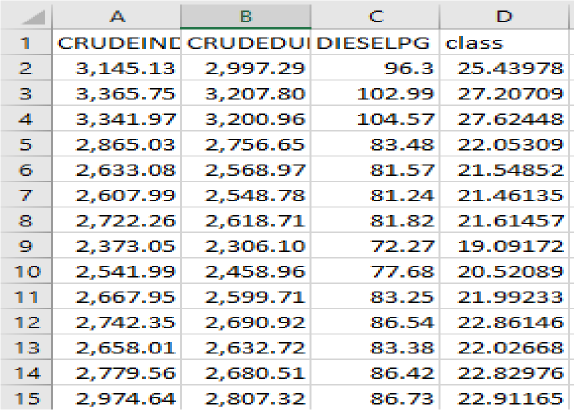
III. DATASET DESCRIPTION

The dataset consists of three attributes which are selected accordingly, As India imports its crude oil from largely from the highest producer of oil dubai.The raw crude oil price is taken into consideration in barrels, the second attribute is taken as to be raw crude oil price in India per barrel. The third attribute is taken as diesel price in India per gallon. Dataset consists of 144 tuples where training set is 100 tuples and 44 tuples are taken as testing set.

(1 Barrel = 159 litres)   (1 Gallon = 3.8 litres)

Feature list (Attributes)

1) Crude Oil price of Dubai (INR)
2)Crude Oil price of India (INR)
3)Diesel price per Gallon (INR)



	A	B	C	D
1	CRUDEIND	CRUDEDUI	DIESELPG	class
2	3,145.13	2,997.29	96.3	25.43978
3	3,365.75	3,207.80	102.99	27.20709
4	3,341.97	3,200.96	104.57	27.62448
5	2,865.03	2,756.65	83.48	22.05309
6	2,633.08	2,568.97	81.57	21.54852
7	2,607.99	2,548.78	81.24	21.46135
8	2,722.26	2,618.71	81.82	21.61457
9	2,373.05	2,306.10	72.27	19.09172
10	2,541.99	2,458.96	77.68	20.52089
11	2,667.95	2,599.71	83.25	21.99233
12	2,742.35	2,690.92	86.54	22.86146
13	2,658.01	2,632.72	83.38	22.02668
14	2,779.56	2,680.51	86.42	22.82976
15	2,974.64	2,807.32	86.73	22.91165

Fig 1. Attributes prices for modeling (INR)[6]

IV. RESULTS AND ANALYSIS

The dataset is divided into test set and training set which is divided as (70% - Training set) (30% Test set).

Model Predicted data point

Original data point

After dividing the dataset into test data and training data we have fitted the model firstly to the Support Vector Regressor which has 3 kernels. They are linear kernel, Polynomial kernel, RBF kernel, Random Forest Regression and Linear Regression.

TEST CASE(A): [1]Prediction using SVR Linear Kernel

Prediction using SVR kernel (linear)
Training accuracy : 0
True Accuracy : 0

Plotting results using Matplotlib:

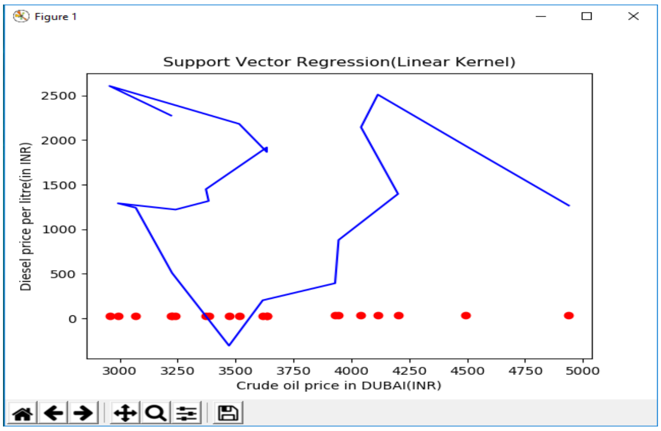


Fig. 2 Represents the graph plotted (Crude oil price in Dubai per Barrel VS Diesel Price in India per litre)

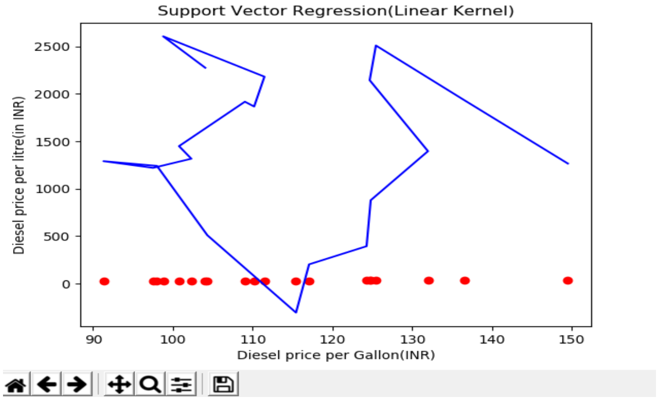


Fig 3. Represents the graph plotted (Diesel price in india per Gallon VS Diesel Price in India per litre).

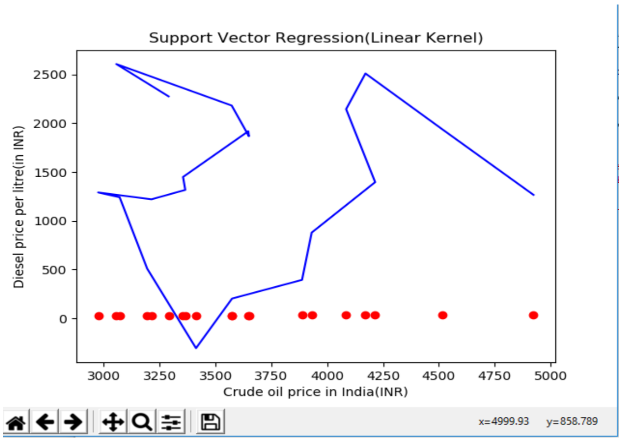


Fig 4. Represents the graph plotted (Crude oil price in India per Barrel VS Diesel Price in India per litre).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The above three graphs are plotted accordingly with the results acquired by Support Vector Regression Linear Model. Hence the values predicted by the model do not touch the original values. Hence the SVR (linear) is not considered to be an accurate model for the prediction for this particular dataset. As, we can see the figures shown above the datapoints are in the range of 30-40 on the y-axis the model has predicted the values above it till 2500 which is deviating from the original points and from this we can say that the SVR(Linear Kernel) cannot predict the price of the oil with the set of datapoints given because of high variance between the predicted points and original points and hence the accuracy for training data is 0(zero) and for the unseen data is 0.

TEST CASE (B):Prediction using SVR (RBF Kernel)[2]

Variance of predicted values : 5.048709793414476e- 29
Variance of Test : 15. 399008063717383
Prediction using SVR Kernel (linear)
Training accuracy :87.5%
True Accuracy : 20%

Plotting results using Matplotlib

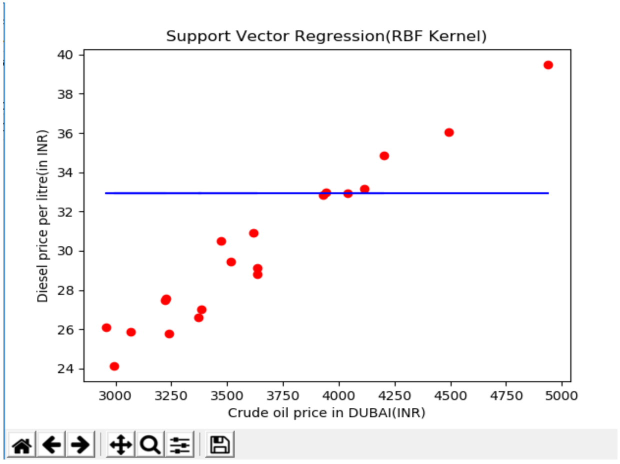


Fig 5. Represents the graph plotted (Crude oil price in Dubai per Barrel VS Diesel Price in India per litre).

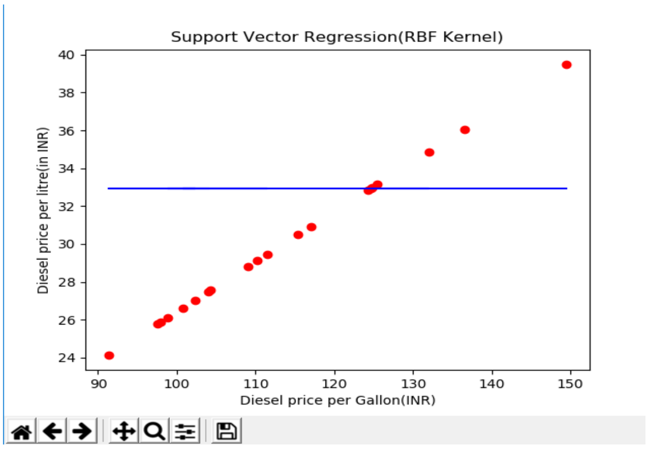


Fig 6. Represents the graph plotted (Diesel price in india per Gallon VS Diesel Price in India per litre).

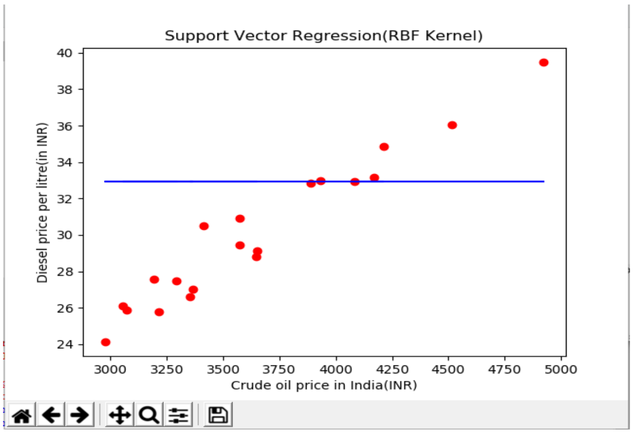


Fig 7. Represents the graph plotted (Crude oil price in India per Barrel VS Diesel Price in India per litre).

The above three graphs are plotted accordingly with the results acquired by Support Vector Regression RBF Model. kernels of SVR rarely scale to the large data points and predict the values. They fit the values to the model but when unseen examplesare given to the model no accurate predictions are given. As a result the training accuracy is 85.7% and test accuracy is 20% from this we can conclude that the rbf function cannot handle this kind of prediction. Generally they try to have a vector to classify the points when it has positive and negative points but in this case it will not be useful.hence this model is rejected for further predictions.

TEST CASE (C): Prediction using SVR poly kernel[2]

Predicting using SVR Kernel (poly)
Training accuracy :0
True accuracy :0

Plotting results using Matplotlib:

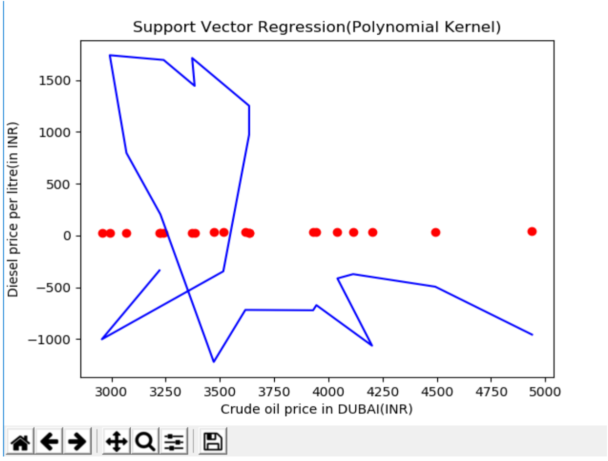


Fig 8. Represents the graph plotted (Crude oil price in Dubai per Barrel VS Diesel Price in India per litre).

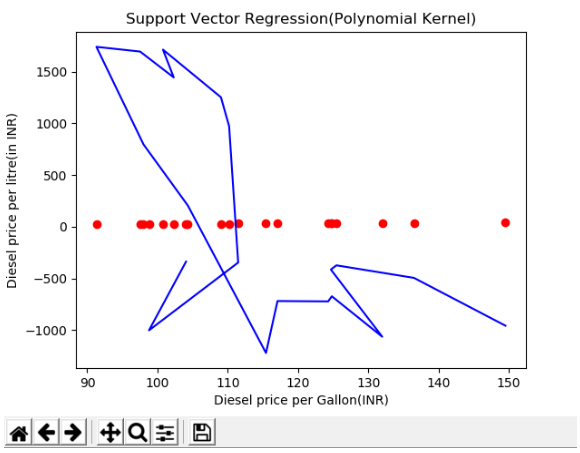
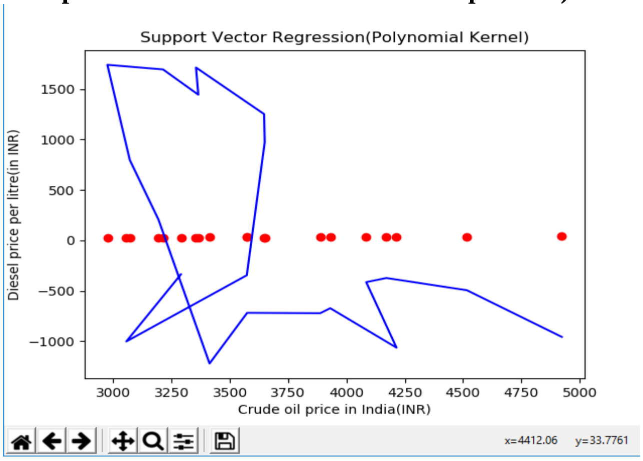

Support Vector Regression(Polynomial Kernel)
Diesel price per litre(in INR)
1500
1000
500
0
−500
−1000
90
100
110
120
130
140
150
Diesel price per Gallon(INR)

Fig. 9. Represents the graph plotted (Diesel price in india per Gallon VS Diesel Price in India per litre).


Support Vector Regression(Polynomial Kernel)
Diesel price per litre(in INR)
1500
1000
500
0
−500
−1000
3000
3250
3500
3750
4000
4250
4500
4750
5000
Crude oil price in India(INR)
x=4412.06 y=33.7761

The Fig. 10. Represents the graph plotted (Crude oil price in India per Barrel VS Diesel Price in India per litre).

The above three graphs are plotted accordingly with the results acquired by Support Vector Regression Polynomial Model. According to the results shown above in the figures, the polynomial kernel of Support Vector Machines have predicted the model which is divergent towards both positive and negative values of Y(Diesel price per litre).As price of object will not be negative we dont need the negative values. Hence the model has predicted the values which have very large variation when compared with the original datapoints. the polynomial equation which bring out the model to fit the datapoint in the polynomial equation, which is not carried out well by the model. Hence the accuracy of the training set provided is 0(zero) and the accuracy of the test set or the unseen datapoints is zero.

TEST CASE (D): Prediction using Random Forest Regressor[3]

Using Random Forest Regressor we have predicted the price of diesel fitting the desired attributes and the following is the source code.

Prediction using Random Forest tree

Training accuracy: 86.66666666666667

True accuracy : 60.0

Plotting the results using Matplotlib

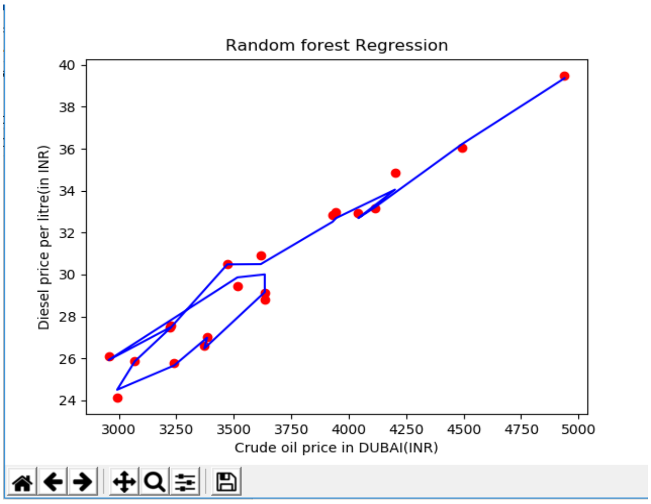
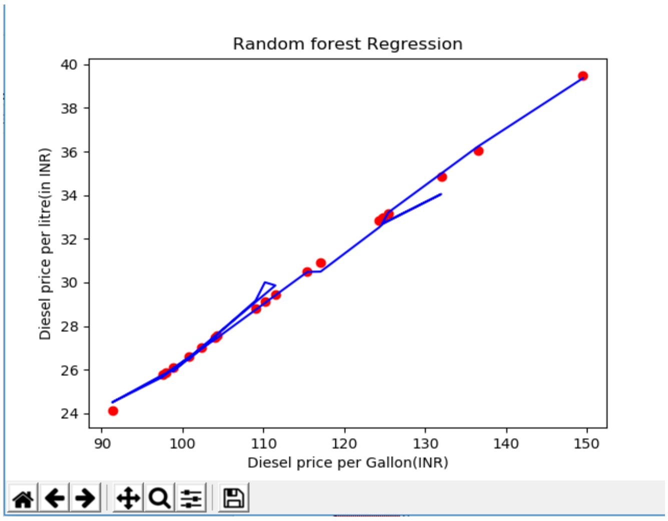


Fig. 11. Represents the graph plotted (Crude oil price in Dubai per Barrel VS Diesel Price in India per litre).



The Fig. 12. Represents the graph plotted (Diesel price in india per Gallon VS Diesel Price in India per litre).

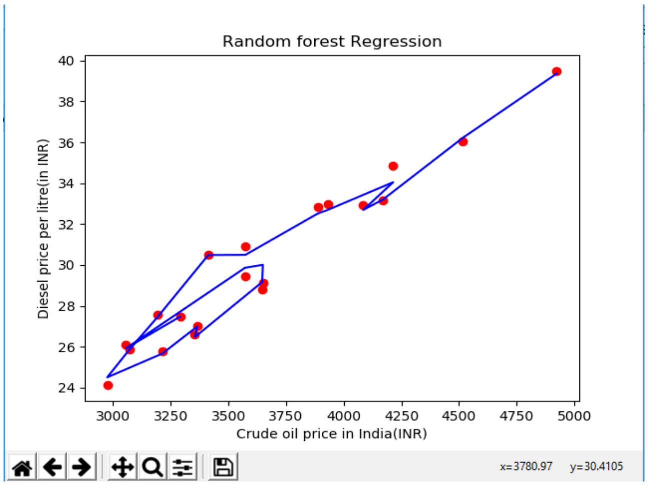


Fig. 13. Represents the graph plotted (Crude oil price in India per Barrel VS Diesel Price in India per litre).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The above three graphs are plotted accordingly with the results acquired by Random Forest Regression Model. As the name suggests the random forest regression is the model which is formed by the cluster of n-decision trees which must be specified in the syntax, more the decision trees less the error. Hence the model random forest has predicted the model with the given datapoints and which when matched with the original datapoints the above are the figures( ). The model has predicted the values very accurately compared with the previous models in SVR. It has predicted with the accuracy of 66.66% on the unseen data points. the accuracy of the training set provided to the model is 80%. Hence this model is said to be accurate.

TEST CASE (E): Prediction using Linear Regression[4]

We have finally used linear regression for predicting the diesel oil price for which we have got accurate results compared to the above models.

Training accuracy: 90.633333333333
True accuracy: 85.0

Plotting results using Matplotlib

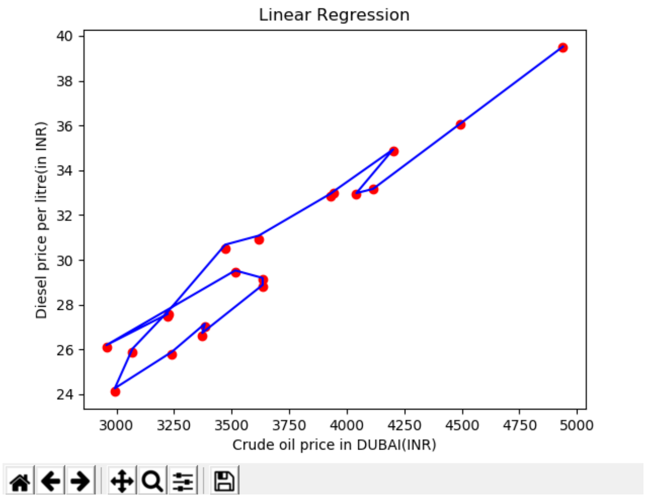
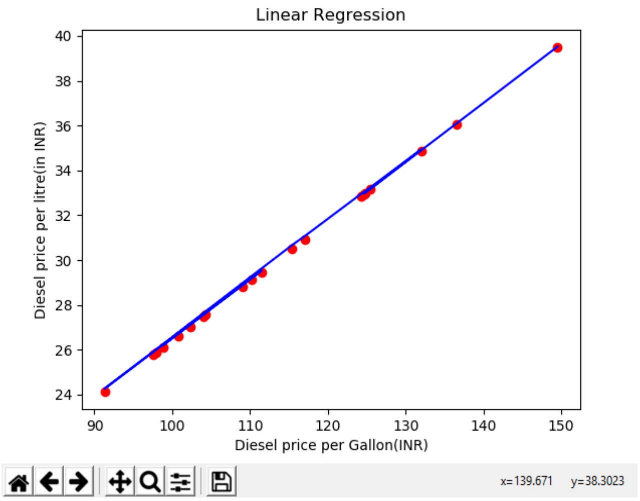


Fig. 14. Represents the graph plotted (Crude oil price in Dubai per Barrel VS Diesel Price in India per litre).



The Fig. 15. Represents the graph plotted (Diesel price in india per Gallon VS Diesel Price in India per litre).

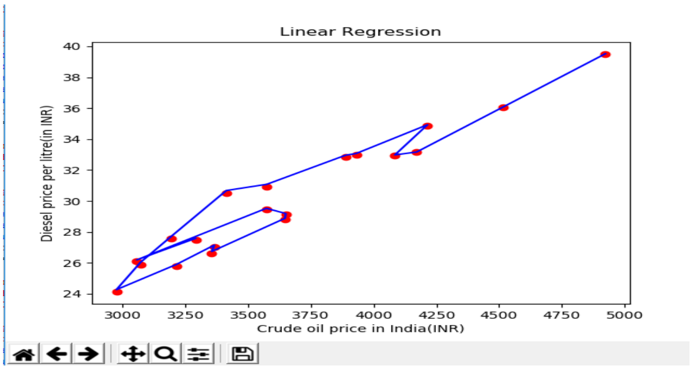


Fig. 16. Represents the graph plotted (Crude oil price in India per Barrel VS Diesel Price in India per litre).

The above three graphs are plotted accordingly with the results acquired by Linear Regression Model. As per the results shown the datapoints supplied to the linear regression algorithm which have been predicted very accurately relative to the previous models. This model accepts the datapoints which are in linear fashion and which are of low variance and predicts them easily.As in the graphs we can see almost all the datapoints are fitted to the model. We have got highest training accuracy of 90.833 % and we have got testing or true accuracy of 85%. Hence, we can use this model for the future predictions of the oil price.

V. CONCLUSION

Forecasting Crude Oil prices is a very challenging problem due to the high volatility of oil prices. In this paper, we developed a new oil price prediction approach using ideas and tools from stream learning, a machine learning paradigm for analysis and inference of continuous flow of non-stationary data. Our stream learning model will be updated whenever new oil price data are available, and provided to model, so the model continuously evolves over time, and can capture the changing pattern of oil prices. In addition, updating the model requires only a small constant time per new data example, The experiment results show that our stream learning model outperformed four other popular oil price prediction models over a variety of forecast time horizons. This process is used to Predict the oil Prices. The prediction model predicts continuous valued functions. To generalize the linear regression model, when dependant variable is categorical and analyzes relationship between multiple independent variables.

ACKNOWLEDGEMENT

We would like to thank our Research Guide Dr. Shaik Subhani, Associate Professor in Information technology, Sreenidhi Institute of Science and Technology, Hyderabad for their continue support and valuable suggestions throughout carried this work. Authors are also grateful to the reviewer for perilously going through the manuscript and giving valuable suggestions for the renovation of manuscript. We would also like to thank the Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad for providing us with the facility for carrying out the simulations.

REFERENCES

1. https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
5. Stock price prediction using machine learning and deep learning techniques by AISHWARYA SINGH (https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/)
6. www.indexwiz.com

AUTHOR'S PROFILE

J Shiva Keerthan, Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Y Nagasai, Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Dr.Subhani Shaik, Associate Professor, Department of IT, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.