Journal of Engineering Sciences
Vol 12, Issue 7, July/2021
ISSN NO:0377-9254

PREDICTING CRUDE OIL PRICES USING MACHINE LEARNING

Dr.CH.RathnaJyothi[1*],Ch.Amulya[2], S.Poojitha[3],N.Varsha[4]
1 Professor, Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, ITI Road, Vijayawada, Andhra Pradesh, India.
2,3,4 Andhra Loyola Institute of Engineering and Technology, ITI Road, Vijayawada, Andhra Pradesh, India
*Corresponding Author's email: chrjyothi@aliet.ac.in

Abstract. Crude oil is the world's most leading fuel. The main advantages of crude oil are it has high density, it is easily available. Oil is used in almost all the industries. Oil is a Constant Power Source. The main aim of this project is to find the different models that efficiently fit the data points and predict the price of fuel with the help of machine learning models. This project works on comparing the different supervised learning models and brings a conclusion based on the efficiency. We have used 3 supervised learning models namely Random Forest Regression, Linear Regression and Decision Tree Regression to know which gives best in terms of accuracy and performance. These algorithms give a numeric value as output. So we can compare the output of these models with the actual models. Now-a-days the oil price has been increasing in leaps and bounds due to certain reasons like inflation throughout the world. Hence these are derived or extracted from petroleum. To predict the values of petroleum like petroleum and Diesel within the future, we've decided to use the Machine Learning algorithms. We use performance metrics to find the performance of the supervised learning models based on their errored value. In this way we can compare different algorithms and find the best one for our problem statement.

Keywords: Prediction; Oil Prices; Machine Learning Models

I. Introduction

Crude oil is the world's most leading fuel. The main advantages of crude oil are it has high density, it is easily available. Oil is used in almost all the industries. Oil is a Constant Power Source. Oil energy is very reliable when compared to other sources such as solar and wind energy. Some machine learning models fit the dataset efficiently depending upon the type of data points provided. The main aim of this project is to find the different models that efficiently fit the data points and predict the price of fuel with the help of machine learning models. This project works on comparing the different supervised learning models and brings a conclusion based on the efficiency. We have used 3 supervised learning models namely, RandomForest Regression, Linear Regression and Decision Tree Regression to know which gives the best in terms of accuracy and performance. These algorithms give a numeric value as output. So we can compare the output of these models with the actual models. Now-a-days the oil price has been increasing in leaps and bounds due to certain reasons like inflation throughout the world. Hence these are derived or extracted from petroleum . The sources of crude oil for India come from neighbouring countries such as Dubai and Saudi-Arabia. To predict the values of petroleum like petroleum and Diesel within the future, we've decided to use the Machine Learning algorithms and apply ensemble learning. Ensemble learning is a technique where we use different algorithms or single algorithms many times. In this way we can compare different algorithms and find the best one for our problem statement.

2 Literature Review

S. N. Abdullah, X. Zeng[1] proposed that among the main factors that affect the volatility of crude oil are the demand and supply of the oil, population and economical aspects. Generalized Autoregressive Conditional Hetroskedasticity(GARCH) model and Naïve Random walk were among the statistical and econometric model used to predict the crude oil price. The models are used to forecaste the crude oil price and then produce a probabilistic prediction for it. The probabilistic prediction is actually generated by running Monte Carlo analysis on annual WTI average prices. Other statistical model predictions made for crude oil price is by C. Morana. This research used semi parametric approach suggested in for short term oil price prediction.

Wei-Yin Loh, University of Wisconsin, Madison, USA[2] proposed that regression learning is a machine learning approach that aims to accurately predict the value

of continuous output variables from certain independent input variables, via automatic estimation of their latent relationship from data. Tree based regression models are

popular in literature due to their flexibility to model higher order non-linearity and great interpretability.

Conventionally, regression tree models are trained in a two stage procedure, i.e, recursive binary partitioning and is employed to produce a tree structure, followed by a pruning process of removing insignificant leaves, with the possibility of assigning multi variant functions to terminal leaves to improve generalization. The primary goal of applying a regressive analysis is usually to obtain precise prediction.

Mr. Brijain R Patel, erMr. Kushik K Rana[4] proposed that reserchers have developed various decision tree algorithms over a period of time with enhancement in performance and ability to handle various types of data . some important algorithms are discussed below. CHID: CHAID(Chi-squared automatic interaction detector) is a fundamental decision tree learning algorithm. It was developed by Gordon V Kass in 1980. CHAID is easy to interpret, easy to handle and can be used for classification and detection of interaction between variables. CHID is an extension of AID(Automatic Interaction Detector) and TIDE(Theta Automatic Interaction Detector) procedures.

Shen Rong, Zhang Bao-wen[7] proposed that linear regression analysis can be divided into simple linear regression and multiple linear regression . It mainly analyses simple linear regression model that is the analysis method of studying the relations between independent variable and dependent variable. To set up linear regression analysis model Python3.6 is used and introduced pandas analysis package and established more advanced data structure and data analysis package of tool.

3 Methodology

Regression analysis is a machine learning approach that aims to accurately predict the value of continuous output variables from certain independent input variables, via automatic estimation of their latent relationship from data.

Algorithms

The algorithms used in this project are as following:

3.1Linear Regression:

Multiple linear regression model will be expressed as followed:

$$y = a_0 + a_1x + a_2x_2+....+e$$

y is the dependent variable and x is the independent variable, a_0 is the constant term, is the intercept of the regression line on the vertical axis and a_1 is the regression coefficient that is the slope of the regression line. e is the random error which will be used to express the effect of random factors on dependent variable [8]. Step wise algorithm is as follows:

STEP 1: IMPORTING LIBRARIES AND LOADING THE DATA.

Import the libraries that might be required to build our model. To get started we imported pandas, Matplotlib, numpy etc.

After importing the libraries, next step will be fetching the dataset and loading our data. The format of the data should be (.csv/.xls).

STEP 2: VISUALISING THE DATA

Visualising the data is important in order to find any correlation between the different parameters.

Matplotlib is excellent library that can be used to visualize our data on various different plots.

STEP 3: FEATURE ENGINEERING

When we visualize our data, we found that there is a strong correlation between the two parameters: date and price. Thereby we will be using these parameters for building our model.

STEP 4: FITTING THE LINEAR REGRESSION MODEL

After that import the method train_test_split from sklearn library. This is used to split our data into training and testing data. Commonly 70–80% of the data is taken as the training dataset while the remaining data constitutes the testing dataset. After that the intercept and coefficient of our model can be calculated.

3.2: Decision Tree Algorithm(For Regression):

A decision tree represents a tree-structured classifier that performs a split test in its internal node and predicts a target class of an example in its leaf node.[9]. Decision trees build regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Consider Fig 1 where X1 and X2 as independent variables and Y as dependent variables. Fig 2 represents the decision tree for the scatter plot. Based on the decision tree, the model made the splits. Now if the new data point lies in between X1=50 and X2=30, it comes under split 4. Now we take the average of all the values in split 4 and assign it to the new data point.

The stepwise algorithm is as follows:

STEP 1: IMPORTING LIBRARIES AND LOADING THE DATA.

Import the libraries that might be required to build our model. To get started we imported pandas, Matplotlib, numpy etc.

After importing the libraries, next step will be fetching the dataset and loading our data. The format of the data should be (.csv/.xls).

STEP 2: VISUALISING THE DATA

Visualising the data is important in order to find any correlation between the different parameters.

Matplotlib is excellent library that can be used to visualize our data on various different plots.

STEP 3: Splitting the dataset into the Training set and Test set

In this step, we have to split the dataset into training set ad test set. We used only 20% of dataset to test the data and remaining 80% of data set used as training set.

STEP 4: Training the Decision Tree Regression model on the training set.

We import the DecisionTreeRegressor class from sklearn.tree and named it as regressor . Then we fit the X_train and the y_train to the model by using the regressor.fit function.

Step 5: Predicting the Results.

we predict the results of the test set with the model trained on the training set values using the regressor.predict function and assign it to y_pred. Step 6: Comparing the Real Values with Predicted Values.

In this step, we compare and display the values of y_test as 'Real Values' and y_pred as 'Predicted Values'.

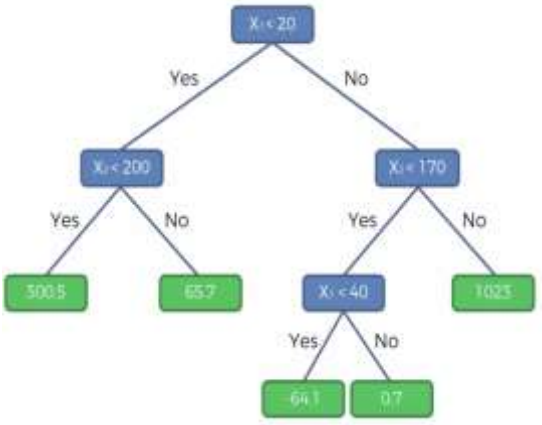

Figure 1: Decision Tree Algorithm

3.3Random Forest Algorithm(For Regression):

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest[10]. Random forest uses ensemble learning. It uses decision trees for n times and predicts the output.

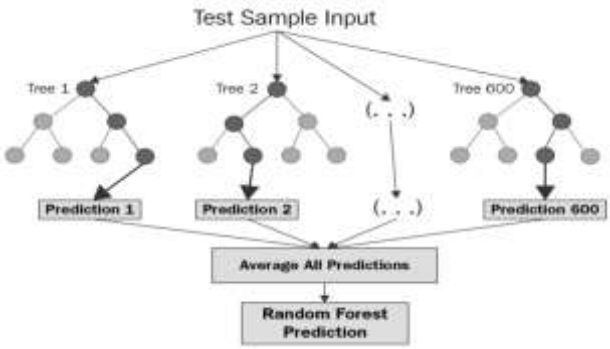

Figure 2: Graph for Random Forest Algorithm

The procedure followed in random forest is as below:

STEP 1 : Firstly, we pick k random data points from the train set.
STEP 2: Then we build the decision tree associated with these k points.
STEP 3: For example, if we want to build 5 decision trees, we repeat step 1 and step 2 for 5 times.
STEP 4 : Each decision tree will predict the output as shown in Fig 2.
STEP 5: For the new data point the average of all the points will be taken as shown in fig 2.

Note : If we consider Fig 2, note that the dataset for all trees will be different(taken randomly).

The generalized algorithmic steps for Random forest algorithm is as follows:

Step 1: IDENTIFIES THE DEPENDENT (Y) AND INDEPENDENT VARIABLES (X)

Dependent variable will be prices while independent variables are the remaining columns left in the dataset.

Step 2: SPLIT THE DATASET INTO THE TRAINING SET AND TEST SET

The training and test split are very important. The training set contains known output from which the model learns off of. The test set then tests the model's predictions based on what it learned from the training set.

Step 3: TRAINING THE RANDOM FOREST REGRESSION MODEL ON THE WHOLE DATASET

From the sklearn package we import the class RandomForestRegressor, create an instance of it, and assign it to a variable. The parameter n_estimators creates n number of trees in your random forest. The .fit() function allows us to train the model, adjusting weights according to the data values in order to achieve better accuracy. After training, then our model is ready to make predictions, which is called by the .predict() method.

Step 4: PREDICTING THE TEST SET RESULTS

Now our random forest model is successfully created.

R^2 score tells us how well our model is fitted to the data

by comparing it to the average line of the dependent variable. If the score is nearer to 1, then it means that our model performs well, if the score is farther from 1, then it means that our model does not perform well.

4Implementation Results

The whole project is based on python, machine learning, and flask.

4.1Data Collection:

We collected the data regarding the crude oil andits pricesfrom Kaggle, Google and Github repositories.

4.2Data Preprocessing:

The raw data cannot be used directly for training the model. Hence we perform preprocessing on the raw data.

First we import the libraries which are frequently used in our project. Here we have used numpy which mainly focuses on operations on arrays, matplotlib for plots and pandas to work on data.

After importing the libraries we import our data set. There may be a chance of missing data in the dataset.Missing data may deviate our results. In order to avoid this we use the SimpleInputer class to replace missing data with mean. Based on the data set and dependent variables we replace missing values will mean, median,constant number etc. If the data set is too heavy and there is only 1% of missing data we ignore the rows with missing values.Now we encode the categorical data using OneHotEncoder class. This method transforms the categorical variable into a set of binary variables (also known as dummy variables). It used N-1 features to show N labels. This improves the machine to understand the data. Next we split the data into a testing set and training set. We apply machine learning algorithms on the training set.In our project we used random forest regression, decision tree regression, simple linear regression. We give the train set as input for the models and train them.

4.3User Interface and output:

A user interface improves the usage of the model. We use flask and HTML to build the user interface.From the above metrics results, we conclude that Random Forest Regression has less error. Hence we use Random Forest Regression in our project. The predicted value will be in Dollar/BBL units as well as in rupees.

Figure 3: User Interface of giving Input

Figure 4: User Interface of getting output

5 Performance Metrics:

In order to evaluate the best model, we useperformance metrics namely, root mean square error, mean absolute error and R2 score. Based on these error metric values, we pick the best model. Those error metrics for regression are as follows :

1 Root Mean Square Error (RMSE)

The RMSE can be calculated as follows:

RMSE = sqrt(1 / N * (sum for i to N (exp_i – pred_i)^2)

Where exp_i is the i'th expected value in the dataset, pred_i is the i'th predicted value, and sqrt() is the square root function

We can say that: **RMSE = sqrt(MSE)**

RMSE for different models used in our project

Linear Regression : 19.3533662859143

DecisionTreeRegression : 9.14731404717267

Random Forest Regression : 1.45273446145987

A perfect RMSE value is 0.0, which suggests that each one's predictions matched the expected values exactly.

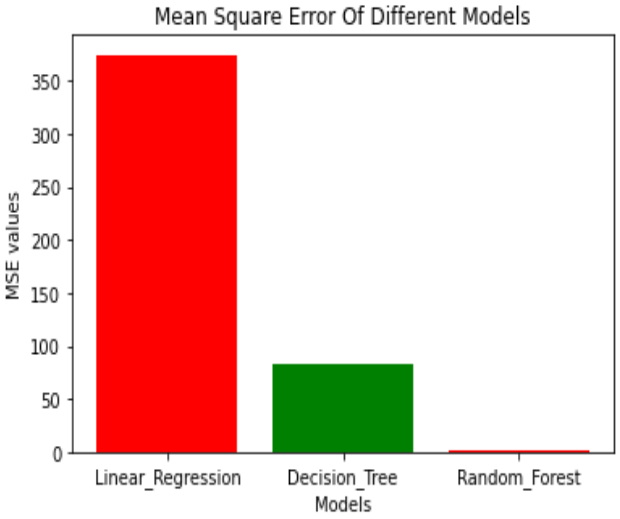

Figure5: Graph for RMSE

2 Mean Absolute Error (MAE)

The average of the absolute error values is called mean absolute error.

MAE = 1 / N * (sum for i to N abs(exp_i – pred_i)),

Where exp_i is the i'th expected value in the dataset, pred_i is the i'th predicted value and abs() is the absolute function.A perfect mean absolute error value is 0.0, which suggests that all predictions match with the expectedvalues exactly.

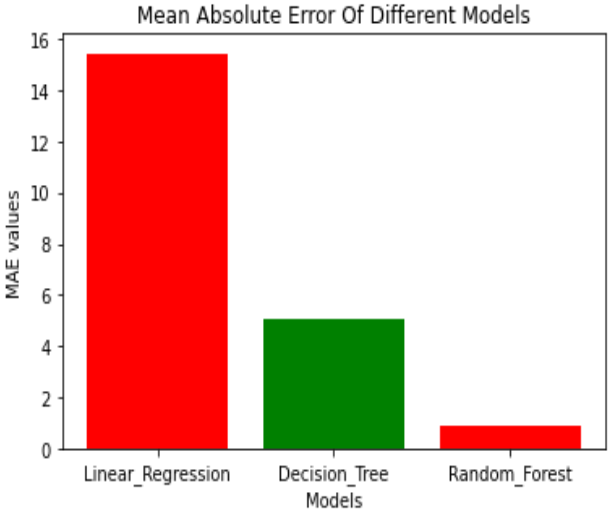

Figure 6 : Graph for MAE

3 R2 Score - Coefficient of Determination

It is the quantity of the variation within the output dependent attribute which is predictable from the input independent variable(s).

The best possible score is 1 which is obtained when the anticipated values are an equivalent because of the actual values.

R2 Score for different models used in our project :

Linear Regression : 0.57380165610804

Decision Tree Regression : 0.90478926790260

Random Forest Regression : 0.997598562970117

Regression	RMSE	MAE	R2-SCORE
Linear	**12.08626473**	**0.5738016561**	**7.6429788**
Decision	**9.147314047**	**0.9047892679**	**5.03809031**
Random	**1.452734461**	**0.99798563**	**0.8729176886**

Table1: Summarized values of performance metrics

6 Conclusion

Machine learning is one of the techniques of Artificial Intelligence which is used for extracting valuable knowledge from large databases[12]. Among all the models used, Random Forest Regression gave us the best results. The error is very less when compared to other regression models.

The proposed system can accurately predict the prices of crude oil which helps us to buy crude oil in advance and decrease the expenses spent.

References

[1]. S. N. Abdullah, X. ZengMachine learning approach for crude oil priceprediction with Artificial Neural Networks-Quantitative (ANN-Q)model.

[2]. Wei-Yin Loh, University of Wisconsin, Madison, USA, Classificationand Regression Trees.

[3]. Lingjian Yang, SongsongLiu, SophiaTsoka, LazarosG.Papageorgiou,A regression tree approach using mathematical programming.

[4]. Mr. Brijain R Patel, 2Mr. Kushik K Rana 1Department of computerengineering, GEC Modasa, India 2Assistant Professor, Departmentof computer engineering, GEC Modasa, India, A Survey on DecisionTree Algorithm For Classification.
[5]. Faliang Huang, GuoqingXie, Ruliang Xiao, Research on EnsembleLearning
[6]. Xu Ying, Ensemble Learning.
[7]. Shen Rong, Zhang Bao-wen, School of information engineering ofNingxia University, YIN Chuan, China,750021, School ofmathematics and statistics,YIN Chuan,China,750021, The researchof regression model in machine learning field.
[8]. Xiangxiang Zeng, Sisi Yuan, You Li, and Quan Zou,Decision TreeClassification Model for Popularity Forecast of Chinese Colleges.
[9]. Leo Breiman Statistics Department University of California Berkeley,CA 94720 January 2001.
[10]. Debasish Basak, Srimanta Pal and Dipak Chandra Patranabis,Support Vector Regression.
[11]. AnnaramSoujanyaa, O. Subhash ChanderGoudb , Sai Prasad. Kc, G Prabhakar Reddy, P.Srinivas Reddy, Featured Based PatternAnalysis using Machine Learning and Artificial IntelligenceTechniques for Multiple Featured Dataset " Elsevier ICAAMM-2016, volume 4 Issue No:88\ ISSN No: 8827–8836 DEC 2017.
[12]. K. Nirosha, B. Durga Sri, Sheikh Gouse and S. Laxmi, Detection ofImage Classifiers Using CNN in Machine Learning.
[13]. Nirosha K., Durga Sri B., Gouse S. Smart heartbeat monitoringsystem using machine learning.
[14]. Sowmya G., Navya K., Divya Jyothi G. Machine learning and miningfor social media analytics.
[15]. AP Gopi (2021), Secure Communication in Internet of Things Based on Packet Analysis, Machine Intelligence and Soft Computing, 2021.
[16]. A Naresh, PG Arepalli (2021), Traffic Analysis Using IoT for Improving Secured Communication, Innovations in the Industrial Internet of Things (IIoT), 2021.
[17]. Bharathi C R, (2017),"Identity Based Cryptography for Mobile ad hoc Networks", Journal of Theoretical and Applied Information Technology, Vol.95, Issue.5, pp.1173-1181.
[18]. Xiangxiang Zeng, Sisi Yuan, You Li, and Quan Zou,Decision TreeClassification Model for Popularity Forecast of Chinese Colleges.