Crude Oil Pride Prediction

Prepared By
Kavya Venkatesh
MEng Department of Computer Science

Abstract

The price of crude oil changes everyday with time and the change depends on factors such as supply and demand and environmental factors. The prediction of the price of crude oil is an essential task for economists and investors who choose to make profits in the oil market. Since the price depends on influential factors these values can increase or decrease at any time with even huge leaps from the previous day's value. Regression can be used to train a model and predict continuous values by setting appropriate values. In this project, we attempt to implement the support vector regression model which has been proved to show more accurate results in terms of price prediction. We build the model and produce results which are more accurate and has lesser errors, in other words the predicted price values are off by only a few values.

1 Statement of the problem

The purpose of this paper is to establish the results that the authors of [1] arrived at. The paper states that support vector regressor provides the best predictive results as comapred to other forecasting models. This paper shows how such a highly accurate support vector regressor model can be developed and shows the results obtained form such a model. The paper also shows results from other types of models by implementing other forecasting models thereby confirming the observation made in the paper[1]. Crude oil is one of the valuable resources in the planet that has a huge demand in the market. It is natural, non renewable resource that serves as a mandatory requirement for the functioning of a lof of industries. Crude oil is considered as a solid investment stock by industrialists and economists. However, the price of this crude oil is fluctuated by several factors such as supply and demand in the market, global temperature, etc.,. Many economists have come up several models to predict and determine the price of crude oil in order to help make investors make the right decision in buying and selling their stocks. These models involve time series analysis, regression based on the history of the market, bayesian network to preditc future prices. Out of these models, based on the paper produced by [1], the regression model developed using support vector machines has proved to be the best. This project is an attempt at showing the efficiency of the support vector regressor model through results produced from such a model which are as accurate to the actual prices as possible. The paper also compares the results produced by other types of models and shows the difference in the results obtained by all the different models.

2 Analysis and Results

2.1 Building the model

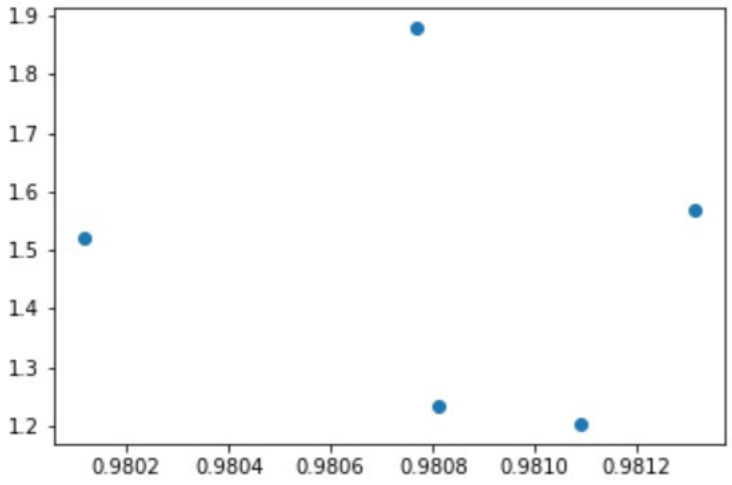
The support vector machines used for classifcation of data, provides importance to misclassifications in its prediction by setting a penalty value to every wrong classifcation.This value is usually set high

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

inorder to obtain high efficieny from the model.The model used in this paper is a support vector machine that is used to perform regression on a series of values.Here, the misclassification value has to be set low in order to accomodate continuous values. The margin vectors in the model plays an imortant role in the accurcay of the values generated as results by the model.The margin will be small for larger C values as the model will correctly classify the points. The larger the C value, the more it enforces the accuracy of the model.The smaller the value, the higher is the probability for misclassification by the model; in this case predicting values that are off by a huge difference.

The gamma value in the support vector regressor is required since the multidimensional hyperplane is not linear. In other words, when the data is not linearly separable, a curvature measure is required to determine the coverage of the radius. A high value of gamma means the curavture is high and a lower value means the area of the hyperplane is very small. The gamma value determines the influence that each sample has on the plane. The higher the value the shorter is the influence range of the samples. This correlation can be better understood by taking the k value from k-nearest neighbor algorithm, where inverse to the gamma in rbf kernel svm, the k value determines how many neighboring samples can be considered.

The choosing of the optimum regularization parameter C and gamma value comes from a series of trial and errors where the model is designed using various pairs of C and gamma values and their performance was analyzed. The performance was measured by two factors; one the confidence on the model to predict the most accurate values for future and two, the root mean squared error of the data which shows the error rate of the predicted values. Figure 1 shows the scatter plot between the confidence and RMSE values of the various versions of the model.



From the figure, we can see that the root mean squared error is less when the confidence of the model is around 98.08 to 98.13 percentage. When the model reaches a confidence level higher than this the model leads to overfitting and the predicted values start to show a slightly increased difference in the true value and the predicted value. The confidence of the model shows how much the model can be trusted based on the predictions that it produces. It shows the accuracy level reached by the model.

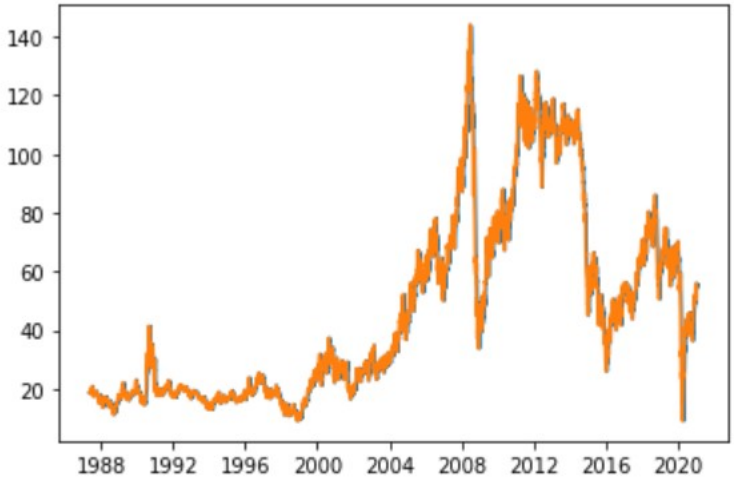
2.2 Data processing

The dataset used in this project is the Brent Oil Prices data that holds the record of crude oil price of each day starting from the 20th of May 1987 to the 25th of January 2021, which is a period of 34 years and has the data for every day that the market was open. The dataset has only two fields namely, the date and the price of crude oil on the given date. Generally on inspecting the domain of crude oil price fixing, it is noted that the two main factors that influence the price of oil are supply and demand of the product in the market and the reviews and reports provided by economists. The latter is used in text based analysis and prediction of how the crude oil market may be influenced

Table 1: Results from ADF test

Parameter	Value
statistics	-2.1199498309908464
p-value	0.23656870472361696
critical value 1%	-3.4311173944575892
critical value 5%	-2.861879136234893
critical value 10%	-2.5669505173087894

which is researched and presented by previous authors. The former is the factor that is taken into consideration in this paper. The supply and demand of the oil in the market decides a kind of price known as the spot price for each type of crude oil. The dataset which is used in the paper is the spot price of crude oil for 34 years. Another reason why the spot price is used because, this price is used as a standard in the market for future analysis as a benchmark for the price on that given day. The spot price is the price that investors base their predictions on for the future value of the oil market. If they find the sopt price to be on a decreasing slope then the investors would predict the market to be on a decline in the future. Similarly, if they notice the market to be on a rise in prices then the investors expect an increase in the profits of their oil investments. A general view of the data is given below.



2.3 Training and evaluation

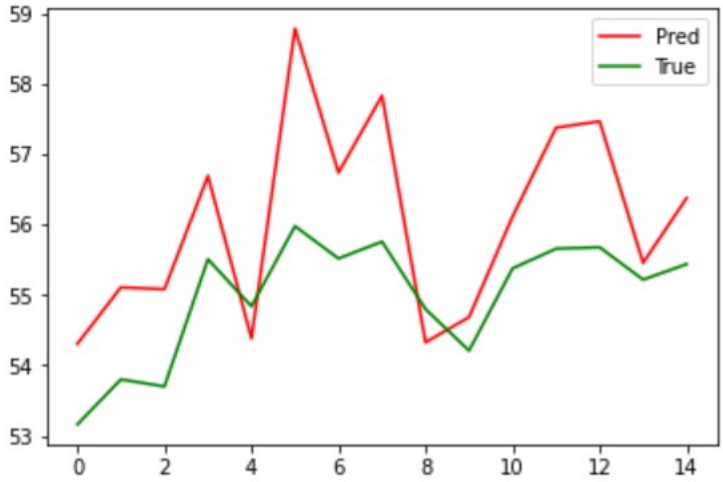
The price field in the dataset is a non-stationary data field. This means that the field type keeps changing with respect to time. In the domain of price prediction that changes on a fixed basis, in this case eveyday it is important to identify the fields that are stationary and non-stationary so that they can be given importance in model training accordingly. Other regression models such as house price prediction, etc., have different types of fields that contribute in the increase or decrease of the value but these fields do not necessarily change with time although they affect the price. In our model we have only one such field which we put under the augmented dickey fuller test to test for stationaryness of the field. With the help of the Augmented dickey fuller test, we determine if the time series generated is stationary or not. The idea behind the dickey fuller test is to find if the seires has an unit root in its expression. We use the default packages available from the statsmodels package to prove the time series of crude oil prices is non-stationary. The values resulting from the dickey fuller test is given in the table 1.

From the results obtained from the test, we can see that the p-value is greater than the 0.05 mark and hence the null hypothesis that determines the presence of an unit root in a series cannot be

rejected.This claim of null hypothesis is that the series where it cannot be rejected is absolutely non-stationary, meaning it can change with time.

2.4 Interpretation of Results

The svm model developed on the dataset produces predicts prices with a confidence level of about 98%. In comparison with other time series models such as ARIMA, the support vector regressor model performed better when the fields are non stationary. Given below are the results from the svm model that was built to predict the oil prices based on data from Brent Oil Prices.



From the figure given above, we can see that the predicted value is very close to the actual value. The predicted price and the actual price plot lines are similar in their growth pattern. The error value is measured by the root mean squared of the predicted values, which equates to a 1.52 approximately. This confidence rate and rms value both are found to be significantly lesser than the ones obtained from ARIMA models as shown by the authors Xie, Wen, et al.[1].

3 References

[1] Xie, Wen, et al. "A new method for crude oil price forecasting based on support vector machines." International conference on computational science. Springer, Berlin, Heidelberg, 2006.

[2] Khashman, Adnan, and Nnamdi I. Nwulu. "Intelligent prediction of crude oil price using Support Vector Machines." 2011 IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2011.

[3] Ahmed, Rana Abdullah, and Ani Bin Shabri. "Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and support vector machines." American Journal of Applied Sciences 11.3 (2014): 425.