Crude Oil Price Forecasting Using XGBoost

Mesut Gumus
Kuveyt Turk Participation Bank
Istanbul, Turkey
Department of Computer Engineering
Selcuk University
Konya, Turkey
mesut_gumus@kuveytturk.com.tr

Mustafa S. Kiran
Department of Computer Engineering
Selcuk University
Konya, Turkey
mskiran@selcuk.edu.tr

***Abstract*—One of the most important role of economic variables in today's world countries are the price and the change of the price of crude oil. Changes in the price of crude oil have a very critical role in terms of treasury and budget, both in company and state planning. For example, one may choose one of the energy or natural gas indexed energy production plans based on the trend of the crude oil price, for planning to meet the need for electricity next year. Accurate forecasting of the crude oil price and realization of the forecasts based on this forecast will provide savings or gains in government and corporate economies, which can reach billions of dollars. There is a great need for this estimation in countries where crude oil production is low and heavily dependent on crude oil import. In this paper, the parameters which are the factors affecting the crude oil prices will be interpreted using XGBoost, a gradient boosting model, from machine learning libraries and estimation will be made.**

Keywords—crude oil; forecasting; gradient boosting machine learning; xgboost

I. INTRODUCTION

Crude oil, which has a very important position in the world economy and is accepted as "black gold", is used in a wide range of fields including industry, transportation, automobile, cosmetics, energy, beginning to be used from the fourth century on, oil has become a turning point in terms of country economies. Crude oil prices are highly sensitive to political and economic developments. Increasing geopolitical risks in the Middle East cause high volatility in oil prices. In addition, the degree of importance that affects prices in countries such as oil reserve amounts, development in economies, global climate changes, changes in supply and demand balance, energy demand are high developments. The Organization of Petroleum Generating Agencies The practices that influence the supply of petroleum by increasing or decreasing production between OPEC member states are also influential on Crude Oil.

These changes in the price of Crude Oil for countries that import oil have an impact in line with their dependence on the country's budget. It is of utmost importance to estimate the Crude Oil price for countries in this situation and to orientate alternative investments if necessary for energy needs.

978-1-5386-0930-9/17/$31.00 ©2017 IEEE

II. DEFINITIONS

A. Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning works on the implementation of software that can be modified when exposed to new instructions [1]. The machine learning process is similar to the data mining process. Both systems search through data to look for patterns. However, instead of extracting data for human binding, the machine uses this data to learn, to detect patterns in data, and to adjust program actions accordingly for data mining software. Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets [2][3].

B. Gradient Boosting

Gradient is a machine learning technique for healing, regression, and classification problems and produces a prediction model of poor predictive models, usually a set of decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [4].

Like other strengthening methods, the degradation boost combines the weak "trainee" into a single strong trainee in a recurrent fashion. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model F to predict values in the form $\hat{y} = F(x)$, by minimizing the mean squared error $(\hat{y} - y)^2$ to the true values y (averaged over some training set) [5].

C. XGBoost

In recent years, machine learning has been generating a lot of curiosity for its profitable application to trading. Numerous machine learning models like Linear/Logistic regression, Support Vector Machines, Neural Networks, Tree-based

models etc. are being tried and applied in an attempt to analyze and forecast the markets. Researchers have found that some models have more success rate compared to other machine learning models. eXtreme Gradient Boosting also called XGBoost is one such machine learning model that has received rave from the machine learning practitioners.

XGBoost is the short name for "Extreme Gradient Boosting" proposed by Friedman in the Greedy Function Approximation: Gradient Boosting Machine journal of the term "Gradient Boosting". XGBoost is based on this original model [6].

The GBM (gradient boosting machine - boosted trees) has been around for really a while, and there are a lot of materials on the topic. This training is working to explain the enhanced trees in a self-contained and primitive manner using controlled learning items. We think this explanation is cleaner, more formal, and motivates the variant used in XGBoost.

It is used for supervised learning problems, on the training data (with multiple features) $x_i x_i$ to predict a target variant $y_i y_i$. Before we dive into trees, let us start by reviewing the basic elements in supervised learning. it is developed with both deep considerations in terms of systems optimization and principles in machine learning [7]. The goal of this library is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate library.

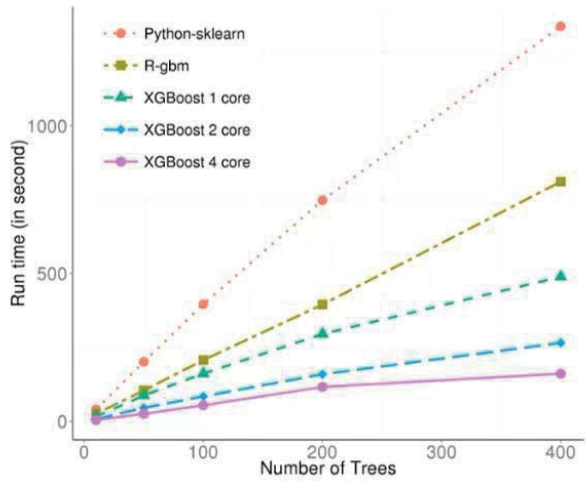

Fig. 1. Performance comparison of XGBoost with R and Sklearn.

D. Crude Oil

It comes naturally. It is an unrefined petroleum product in hydrocarbon deposits and contains of other organic substances. Crude oil can be refined to produce usable products such as gasoline, diesel and various forms of petrochemicals. This resource, also known as fossil fuel, cannot be renewed. This naturally means that we cannot change it in the way we consume it, and for this reason it is a limited resource [8].

Crude oil is usually obtained by oil drilling, where it is found in other sources, such as natural gas (which is softer and above this crude oil) and brine (denser and then sink). It is then refined and processed in various forms such as gasoline, kerosene and asphalt and sold to consumers.

Although it is often called "black gold," crude oil has ranging viscosity and can vary in color to various shades of black and yellow depending on its hydrocarbon composition. Distillation, the process by which oil is heated and separated in different components, is the first stage in refining.

III. Material and Method

A. Data Preparation

This is the "crude" data of crude oil price for period from February 2009 to May 2017 in the image below. There was been a significant price jump somewhere in 2014. This free-available data only contains prices for each working day and we will try to make a day-ahead price forecast, which is somehow ill-formed from the beginning. We can't follow price trends during the day.

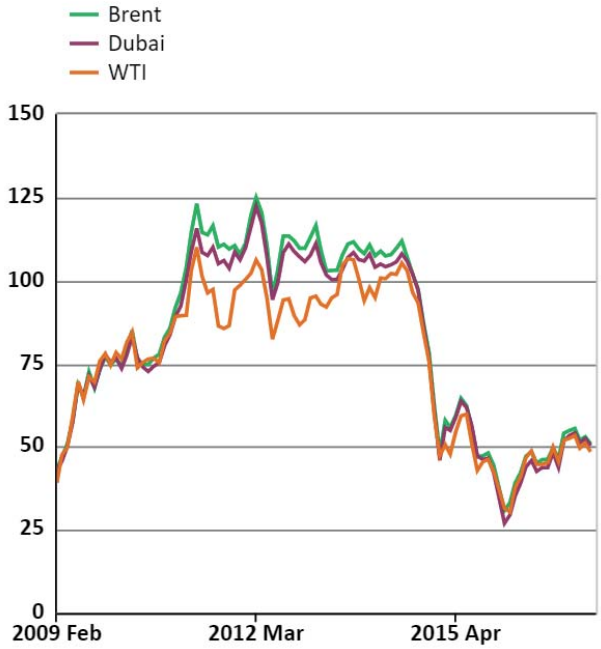

Fig. 2. Crude oil trends.

TABLE I. Monthly Spot Crude Oil Prices

Units	Monthly Spot Crude Oil Prices		
	Brent ($/bbl)	*Dubai ($/bbl)*	*WTI ($/bbl)*
2015 May	64.56	63.69	59.27
2015 Jun	62.34	61.78	59.8
2015 Jul	55.87	56.25	50.9
2015 Aug	46.99	47.22	42.86
2015 Sep	47.24	46.15	45.45

2015 Oct	48.12	46.55	46.2
2015 Nov	44.42	42.22	42.7
2015 Dec	37.72	34.77	37.23
2016 Jan	30.8	27	31.54
2016 Feb	33.2	29.5	30.39
2016 Mar	39.07	35.18	37.77
2016 Apr	42.25	39.04	40.96
2016 May	47.13	43.95	46.73
2016 Jun	48.48	45.83	48.75
2016 Jul	45.07	42.62	44.69
2016 Aug	46.14	43.73	44.75
2016 Sep	46.19	43.74	45.2
2016 Oct	49.73	48.26	49.89
2016 Nov	46.44	43.77	45.57
2016 Dec	54.07	51.78	52.01
2017 Jan	54.89	53.37	52.51
2017 Feb	55.49	54.17	53.4
2017 Mar	51.97	51.16	49.58
2017 Apr	52.98	52.45	51.06
2017 May	50.87	50.31	48.5

I also used the price trends of gold, silver and natural gas in the same period. These prices change prices in line with crude oil prices. As this is only a machine learning-focused practice, there is no need for further data mining at the OPEC meetings to investigate key global political events and other relevant factors that can be used in the related (crude oil) raw materials.

B. XGBoost Configuration

The parameter definitions required for XGBoost are set as follows.

```
eta_list = [0.1, 0.05, 0.04, 0.03, 0.02, 0.01]
min_child_weight_list = [6, 5, 4, 3, 2]
subsample_list = [0.6, 0.7, 0.8, 0.9]
colsample_bytree_list = [0.9, 0.8, 0.7]
max_depth_list = [3, 5, 7, 8, 9]

gb_params = {'objective': 'reg:linear',
             'eval_metric': 'rmse',
             'seed': 2016,
             'eta': 0.037,
             'min_child_weight': 3,
             'subsample': 0.7,
             'colsample_bytree': 0.83,
             'silent': 1,
             'max_depth': 8
             }
```

C. Forecast Proccessing

We use the xgboost function to train the model. The arguments of the xgboost function are shown in the picture below.

```
xgboost(data = NULL, label = NULL, missing = NA, weight = NULL,
  params = list(), nrounds, verbose = 1, print_every_n = 1L,
  early_stopping_rounds = NULL, maximize = NULL, save_period = 0,
  save_name = "xgboost.model", xgb_model = NULL, callbacks = list(), ...)
```

Fig. 3. xgboost() method definition.

The data-independent variable in the Xgboost function is for the input properties data set. Accepts a matrix, dgCMatrix or local data file. The nrounds argument refers to the max number of iterations (i.e. the number of trees added to the model). The argument of obj specifies the customized objective method. Returns the grade given and degrade with dtrain and the second grade.

```
# Train the xgboost model using the "xgboost" function
dtrain = xgb.DMatrix(data = X_train, label = Y_train)
xgModel = xgboost(data = dtrain, nround = 5, objective = "bina
ry:logistic")
```

Fig. 4. XGBoost training method definition.

We can also use the cross-validation function of xgboost i.e. xgb.cv. In this case, the original sample is randomly partitioned into nfold equal size subsamples. For nfold sub-examples, a single sub-sample is stored as verification data to test the model, and the remaining (nfold-1) sub-samples are used as training data. The cross-validation process is then repeated nrounds times, with each of the nfold subsamples used exactly once as the validation data.

```
# Using cross validation
dtrain = xgb.DMatrix(data = X_train, label = Y_train)
cv = xgb.cv(data = dtrain, nround = 10, nfold = 5, objective =
"binary:logistic")
```

Fig. 5. XGBoost validation method definition.

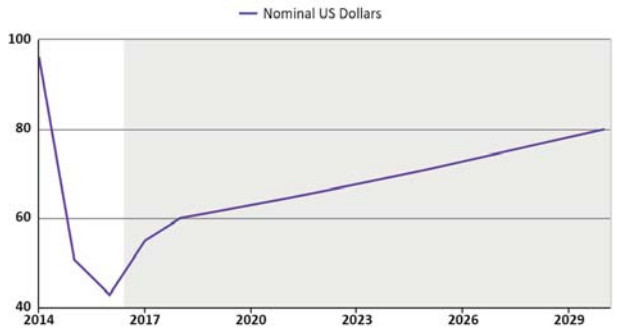

Fig. 6. Crude oil forecast trend.

TABLE II. THE FORECAST RESULTS OF CRUDE OIL

Units	**$/bbl**
Year	
2017	55.2
2018	55.1
2019	54.1
2020	54
2021	54.4
2022	55.2

ACKNOWLEDGMENT

Finally, gradient boosting has proven many times to be an effective prediction algorithm for both classification and regression tasks. By selecting the number of components included in the model, we can easily control the so-called bias-variance trade-off in the estimation. In addition, component wise gradient boosting increases the attractiveness of boosting by adding automatic variable selection during the fitting process.

REFERENCES

[1] A. Munoz, "Machine Learning and Optimization", Courant Institute of Mathematical Sciences, New York, NY.

[2] Margaret Rouse, Machine learning, January 2011

[3] Margaret Rouse, Machine to Machine Learning, June 2010

[4] A. Jain, M. N. Menon, S. Chandra, Sales Forecasting for Retail Chains, 2016

[5] Z. Zhi-Hua, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC. 2012, pp. 23.

[6] T. Chen, T. He, Xgboost: extreme gradient boosting. R package version, 2015, 0.4-2.

[7] Tianqi Chen. Xgboost https://github.com/tqchen/xgboost

[8] J. D. Hamilton, Understanding crude oil prices, National Bureau of Economic Research, 2008