Crude Oil Price Predictor 2020

Christopher Page

8/12/2020

Introduction/Overview/Executive Summary

Oil is an important commodity. Some of our planet's inhabitants produce it. Many more consume it. All are affected by it. Because of its wide-ranging effects, oil commands collective attention enabled by analysis. One of the key metrics to take into account in the analysis of the oil indutry is the daily closing price of crude oil futures traded on a global basis in such venues as the New York Mercantile Exchange.

Methods for predicting that price may entail studying the quantitative and qualitative factors at play inside the oil industry as well as in such adjacent industries as transportation and manufacturing. They may also entail studying the often complex diplomatic, military, economic, cultural, and environmental developments that help explain why crude oil closing prices rise and fall as they do over time.

This project proposes a simpler method, one that predicts crude oil closing prices based on time and the closing prices of complementary (e.g., gasoline) and competing (e.g., platinum) commodities. The premise is that there is much to be learned by studying the behaviors of commodity traders engaged in making daily investment decisions based on running evaluations of risks and rewards.

That premise informed the development of algorithms with the potential to predict crude oil closing prices with a viable level of accuracy, quantified as the Root Means Square Error (RMSE). Using a 25,190-row data set constructed with publicly available information downloaded for free from https://www.nasdaq.com/, the author developed and then evaluated fifteen such algorithms.

The evaluation process led to the selection of one algorithm that produced the lowest RMSE when applied to a randomly generated test set. That algorithm used a Ranger model [1] which took into account the *year* and *month* components of time as well as the closing prices of four other globally traded commodities: (1) *heating oil*, (2) *gasoline*, (3) *platinum*, and (4) *soybeans*.

The RMSE in question was **2.93**, a figure equating to 12% of the 23.51 produced by an initial algorithm that only took into account the average closing price of crude oil across the period of observation. Because of the relatively small RMSE, the selected algorithm could be considered analytically viable and, therefore, of use to those engaged in analzing the oil industry.

Method/Analysis

The first step was to load all of the necessary *R* libraries from the publicly available repository http://cran.us.r-project.org. Those libraries included *caret*, *caTools*, *dplyr*, *forcats*, *foreach*, *gam*, *ggplot2*, *ggrepel*, *ggthemes*, *knitr*, *lubridate*, *purrr*, *randomForest*, *ranger*, and *tidyverse*. The *randomForest* and *ranger* packages proved to be of particular value because of the models they introduced.

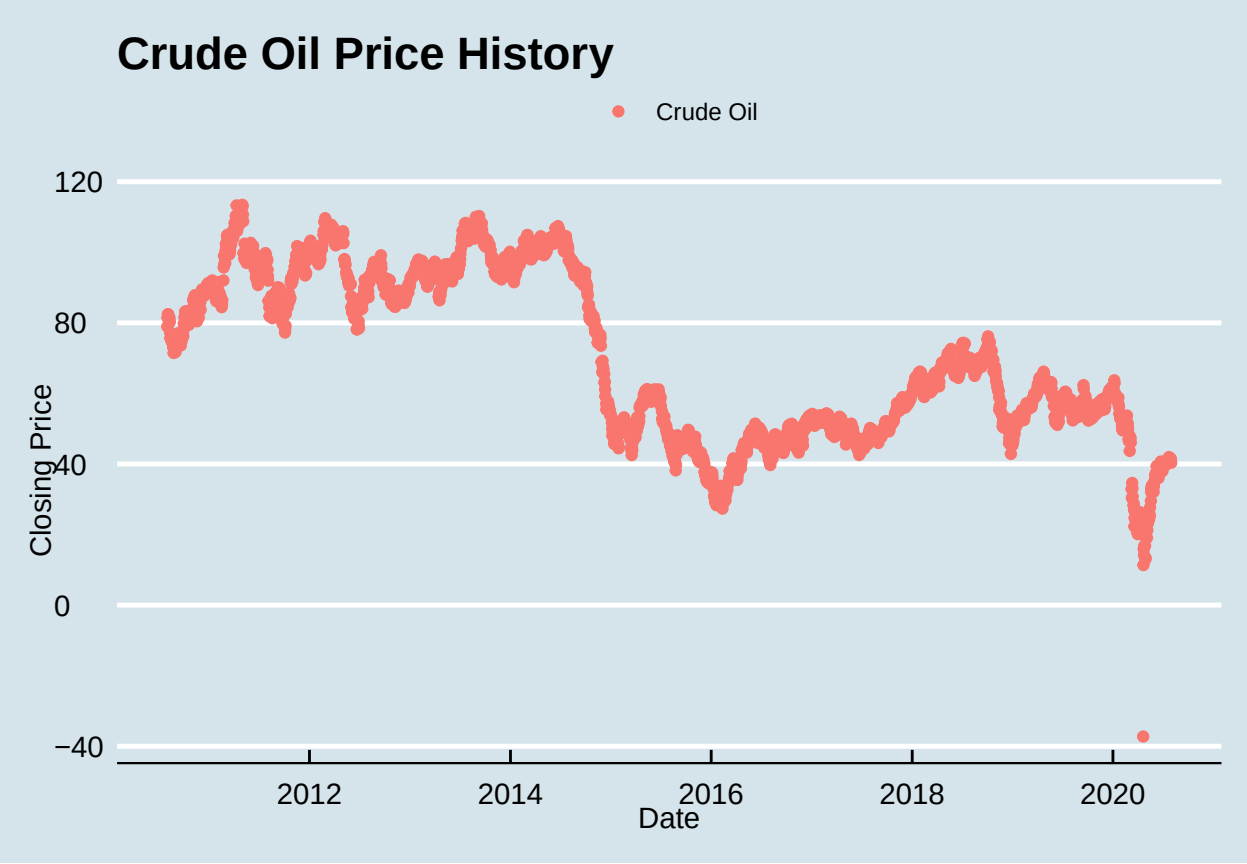
The second step entailed loading *commodity_closing_prices_2010_2020.csv*, a file containing commodity closing prices for a period of observation extending from the beginning of August 2010 to the end of July 2020, and then wrangling the data in a manner that would result in ready access to information detailing

[1] Characterized by https://www.rdocumentation.org/ as a "fast implementation of random forests"

the cyclical and linear components of time as well as the daily closing prices of crude oil and nine other commodities from the energy, precious metals, and agriculture sectors.

Wrangling enabled explotatory data analyses that began with an examination of crude oil in isolation across the ten-year period of observation. A plot generated through that examination made clear the history of crude oil closing prices as they rose and fell, often sharply, between the end of July 2010 and beginning of August 2020. Of particular note was the rapid descent deeply into negative territory in April 2020 [2] That descent helped account for prices having the following minimum, maximum, mean, and standard deviation.

Parameters	Values
Mininum	-37.25000
Maximum	113.45000
Mean	70.13307
Standard Deviation	23.52209



[2]See https://www.nytimes.com/2020/04/20/business/stock-market-live-trading-coronavirus.html for an example of news reports covering that event, widely characterized as “unprecedented”.