ORIE 5741 Project: Prediction of Oil Prices
Midterm Report

Harshavardhan Bapat (hsb57); Joel Dsouza (jnd74); Yash Ganatra (ybg3)

1. Introduction

We attempt to employ machine learning techniques to build a model to accurately predict crude oil prices using economic and financial data. Initially, our broader goal was to isolate the conditions that lead up to oil crises, but we found it more prudent to primarily focus on accurately predicting oil prices, and then tune our model to attempt to capture the conditions leading up to major price moves. Below we outline the preparation of the data that is considered, our preliminary analysis, and the next steps we can carry out.

2. Data Preparation

Oil prices, like most other commodity and asset prices, are a function of supply and demand. Supply of global crude oil is largely in the hands of a few top producers with access to the largest oil fields in the world – the OPEC coalition, United States of America, Russia, Canada and China. An important factor to consider on the supply side is the cost to store and ship oil barrels, and shipping bottlenecks often influence oil prices. On the other hand, demand for crude oil is tougher to estimate, as crude oil is one of the basic sources of energy currently known to mankind and is a major input for most industrial businesses. We attempt to capture demand for oil by considering proxies for majority of global industrial demand through various economic and financial factors. The data that has been considered can be classified into three categories:

1. Supply Side Factors
2. Demand Side Factors
3. Other Economic and Financial Data

The below table summarizes the sources and frequency of data considered:

Data	**Frequency**	**Data**	**Frequency**
WTI Crude Oil Futures	Daily	OPEC Oil Production per day	Monthly
CASS Freight Rates	Monthly	US Oil Production per day	Monthly
US GDP	Quarterly	Canada Oil Production per day	Monthly
China GDP	Quarterly	Russia daily Oil Production per day	Monthly
Russia GDP	Quarterly	US Inflation YoY	Monthly
Japan GDP	Quarterly	Euro Area Inflation YoY	Monthly
Germany GDP	Quarterly	Natural Gas Futures	Daily
USD/EUR Exchange Rate	Daily	Euro Area Ind Production Index	Monthly
US Ind Production Index	Monthly	Russia Ind Production Index	Monthly

Our dependent variable is the WTI Crude Futures Price, which is the closing price for futures of the West Texas Intermediate Crude Oil delivery contract with nearest maturity. For supply side factors, we consider the daily oil production of OPEC, US, Russia, Canada and China as reported by the DOE, and the CASS Freight Rate Index which is a proxy for shipping costs. To proxy for demand side factors, we use the YoY GDP of some of the largest countries in the world, ensuring we capture distribution across all regions, while we also consider the Industrial production output of US, Russia and the Euro Area, which may be a more concise proxy for crude oil demand, since industrial businesses are the major sources of demand (as opposed to industries such as software etc.). Commodities such as crude oil are also closely related to inflation, but the relationship may

be a two-way causation. In addition to the above data, we also consider 5 major global stock indices as proxies for 'economic health' of the world, namely S&P500 (US), Dow Jones Industrial Average (US), DAX (Germany), Hang Seng (Hong Kong) and Nikkei 225 (Japan). We also consider US 10- year government bond yields as another indicator of economic health, and the USD/EUR exchange rate as a proxy for relative strength of USD and EUR, two of the most prominently used global currencies. All the above data is sourced from Bloomberg LP.

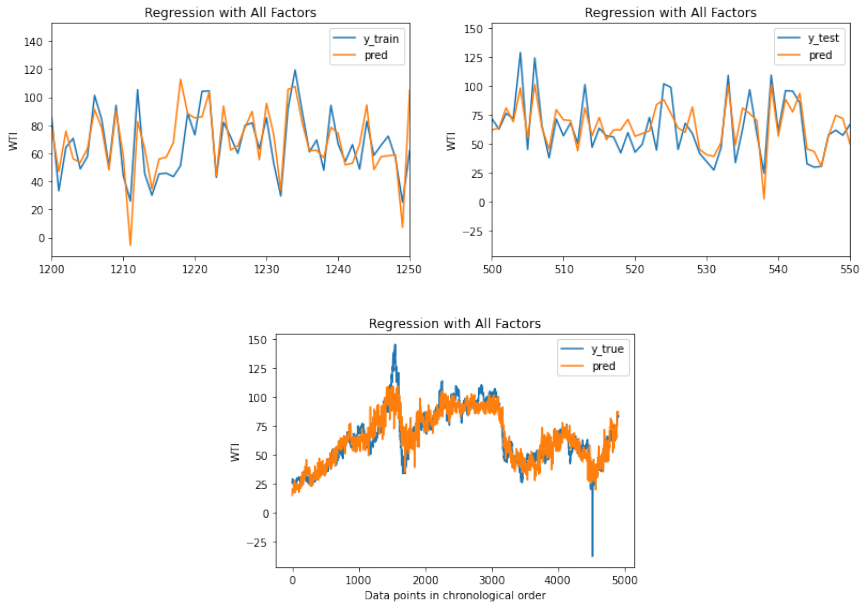
We face 3 main challenges while compiling our feature set. Firstly, each dataset does not extend to 1990 as we would prefer. For now, we drop the missing values, and restrict our model from 2002-2021. Secondly, our data is a combination of data of different frequencies, and since our dependent variable is reported daily, we must backfill the rest of the features based on the latest previously known data. However, along with this, since we are attempting to model a time series, we must take care to avoid forward looking biases. Therefore, our features must be lagged, and in addition, we must ensure we are not considering data that has not been released at the time in consideration. Most economic data is released at a one-month/quarter lag, so we lag our backfilled data by one month/quarter further. To account for the gradual changes in the economy, we also take a monthly/quarterly moving average of these lagged values as per necessary. This is just to ensure we have a proxy for changing environments, to avoid our feature information remaining stagnant. Stock prices and bond yields have significant momentum information, and this may be captured in lagged values. We suspect this information may be significant in contributing to our dependent variable, therefore, we also include 1, 2, and 3 day lagged values in our overall feature set. After the above preprocessing steps, we obtain our overall initial feature set consisting of 61 features, including lagged economic data, lagged oil prices and volumes, lagged supply data, and lagged financial data.

3. Preliminary Analyses

We started by carrying out a preliminary least-squares regression for oil prices with only the oil production data as our features. As expected, we found a negative correlation with US and OPEC oil production, but on the other hand, a positive correlation with Russia and Canada oil production volumes, by looking at the weights. We did not find the model very useful as it predicted the prices with a high root-mean-squared error of around 18.02 and 18.65 for train and test datasets respectively, and the model could not capture the jumps in oil prices accurately. We have used root-mean-squared error instead of MSE to represent the error in the same units as oil price to get a better understanding of model performance. We have added the graphs for the model below, with a zoomed in view to make it easier to view the performance of the models and the fluctuations in prices.

Next, we carried out regression on the full set of our features (excluding the lagged oil price), which resulted in reduction of our RMS error as compared to the preliminary regression. We

observed a positive correlation with freight rates, which is expected since oil prices increase with shipping costs. GDP of the 5 nations in consideration also observed a slight positive correlation, but a stronger correlation was observed with the industrial indices in consideration. We still fail to capture the major jumps in prices using this model.



We are now concerned about overfitting with too many features being used. The next step hence, was to use the ControlBurn algorithm to reduce our number of features. After using ControlBurn, our features reduce to 16, but running a regression on the features increases our RMS error slightly from our full feature regression. The auto-correlation function for oil prices showed that a large number of lagged terms had significant correlation, implying that the oil prices were seasonal. The histogram of prices also signifies that prices are not mean-reverting. We thus noticed that oil prices observe a seasonal trend based on demand, supply, and other macroeconomic features, and believe that it would be a prudent to test our analysis on oil price returns, rather than the oil prices themselves, since returns exhibit are better fitted to a normal distribution.

Consequently we plotted the auto-correlation function for the oil price daily returns, and noticed significance up to the first 3 lagged terms, with some degree of significance in the later terms as well, but we would like to focus our analysis on up to the 3rd lagged term only. This suggested that using lagged features as explained earlier is a prudent choice in our model. It also indicates that using returns as the dependent variable could be a viable option, rather than the oil prices themselves.

4. Next Steps

We constructed our dataset using features from different resources and found out that most of them were correlated. We are aware that multicollinearity may affect our predictions as it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. Therefore, we used ControlBurn to find a set of explainable features but again most of them seemed correlated in theory. Our next step is to run Lasso regression on these features to get a more refined set of features that will work equally well. We will also filter the remaining features using Pearson correlation.

A different option may be to analyses returns rather than prices as returns as normally distributed, but a possible issue we might face with this model is since returns are mean reverting, we may fail to capture the major fluctuations over time. We plan to include some more macroeconomic data and global features which might explain these sudden movements. We also plan to build a model to classify the price movements as UP or DOWN.

Another approach we want to explore is that of reducing the dimensions of our data using PCA. This will help us to capture the variance of the dataset using fewer number of dimensions although we understand that it would reduce the interpretability of the model.