Forecasting Crude Oil Prices

Jose A. Bird

1. Introduction

Estimating the future price of crude oil is critical to plan for exploration of oil and gas properties. The time horizon to develop an oil field could be up to 5 years due to the lead time require for engineering, environmental permitting, facility fabrication and construction activities. An estimate of the future crude price at the time of commercial startup is needed to determine if the project can be economically justified.

In addition to field development activities, integrated oil companies and downstream refining companies purchase significant crude volume quantities to process at their refineries and engage in crude hedging activities to minimize commodity risk. Therefore, integrated oil & gas companies and downstream refining companies have a strong interest in a good estimator of crude oil prices. Midstream oil companies, who transport the crude oil, would also be interested to plan their future transportation logistics expansion projects.

2. Data

The data for this project is available via the St. Louis Fed Bank website (**https://fred.stlouisfed.org/**) which includes US economic data and price data for several commodities. This data is a compilation of several data sources including Bureau of Labor Statistics (BLS), US Energy information Administration (EIA), US Census Bureau, and the Board of Governors of the Federal Reserve System. The data is first collected via an excel add-in and a comma delimited file created to load into a Jupyter notebook via the pandas library. The data considered for this study represents average monthly prices and covers the years 2001 through 2020. The data will be used to build several models to predict crude oil prices. Explanatory variables considered include US unemployment rate, consumer price index (CPI), housing starts, consumer credit, total capacity utilization, automobile sales, and other relevant economic indicators.

3. Methodology

The first step in the study consisted of conducting an explanatory data analysis to examine the distributions of independent and dependent variables and the relationships among these variables. Table 1 provides the summary statistics generated using the pandas describe method. Examination of the maximum statistic reveals a significant outlier in the layoffs column. For purposes of this analysis, layoffs values exceeding 5,000 were replaced with the median value of 1,882 given in the table. Figure 1 shows histogram of the variables before and after this change to the data. Note that crude price has some extreme values with prices exceeding 100 $/bbl.

statistic	Oil_Price	Layoffs	Unemp_rate	CPI	HOUST	PERMIT	New_houses	ALTSALES	Cap_util	Cons_credit	POP	US_stock_change
count	238.00	238.00	238.00	238.00	238.00	238.00	238.00	238.00	238.00	238.00	238.00	238.00
mean	62.36	1969.18	6.08	219.34	1253.72	1303.02	949.42	15.59	76.63	2861.09	309394.32	1353.76
std	25.82	745.56	1.98	24.93	466.56	477.30	285.91	2.27	3.16	705.30	13817.59	12160.61
min	16.55	1437.00	3.50	175.60	478.00	513.00	414.00	8.72	64.24	1729.85	283920.00	-29855.00
25%	42.55	1786.25	4.70	199.15	915.50	977.00	719.25	14.77	75.29	2310.06	297485.00	-7067.50
50%	59.02	1882.00	5.55	220.03	1203.50	1272.00	1023.50	16.41	77.00	2662.32	310677.30	1517.50
75%	83.76	1993.75	7.45	237.81	1628.00	1663.50	1151.00	17.08	78.72	3440.15	321690.91	9044.25
max	133.88	11489.00	14.70	260.33	2273.00	2263.00	1424.00	21.71	81.13	4209.63	330618.50	49569.00

Table 1 Summary Statistics

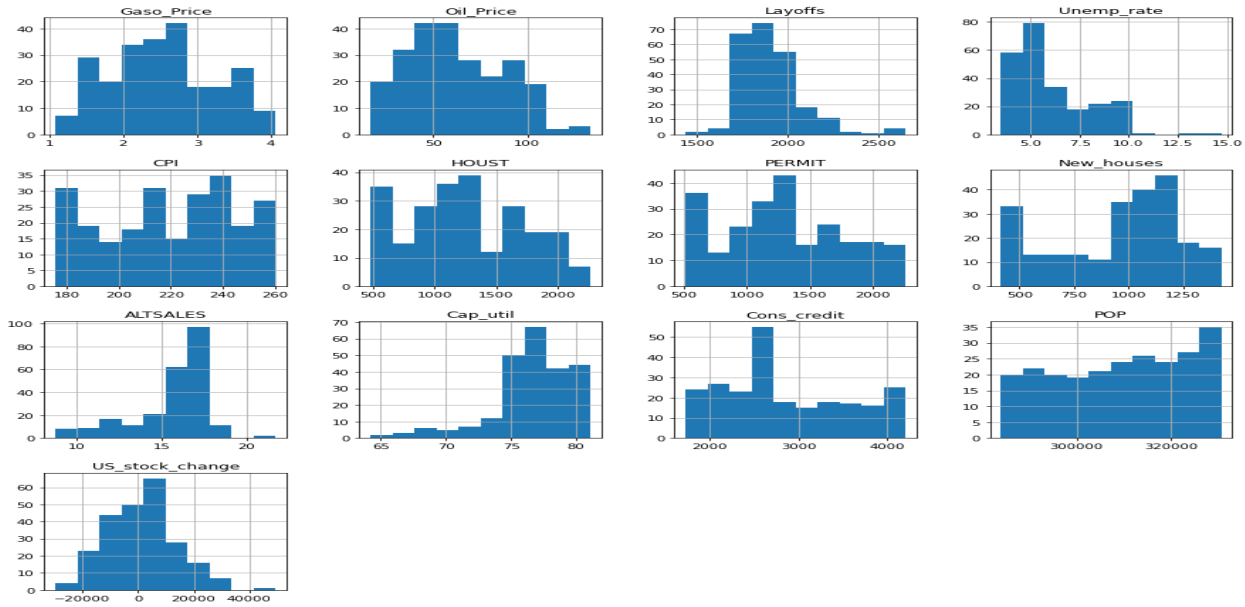

Figure 1 Data Distributions

Figure 2 below shows the correlation matrix represented as a heat map illustrating the very high correlation between crude price and gasoline price. For this analysis, the goal was to obtain a relationship between crude oil price and the economic variables so gasoline price was excluded as a predictor. Note that housing starts, permits, and new private houses have a moderate negative correlation with crude prices which means that crude prices decline when these three variables are high. Other interesting variables that have a moderate positive correlation with crude price were the CPI which is a measure of inflation, the US population size, and the unemployment rate. As a side note, interesting that gasoline price has a higher correlation with population size and the CPI when compared to crude price which suggests a more direct relationship between these variables and gasoline price.

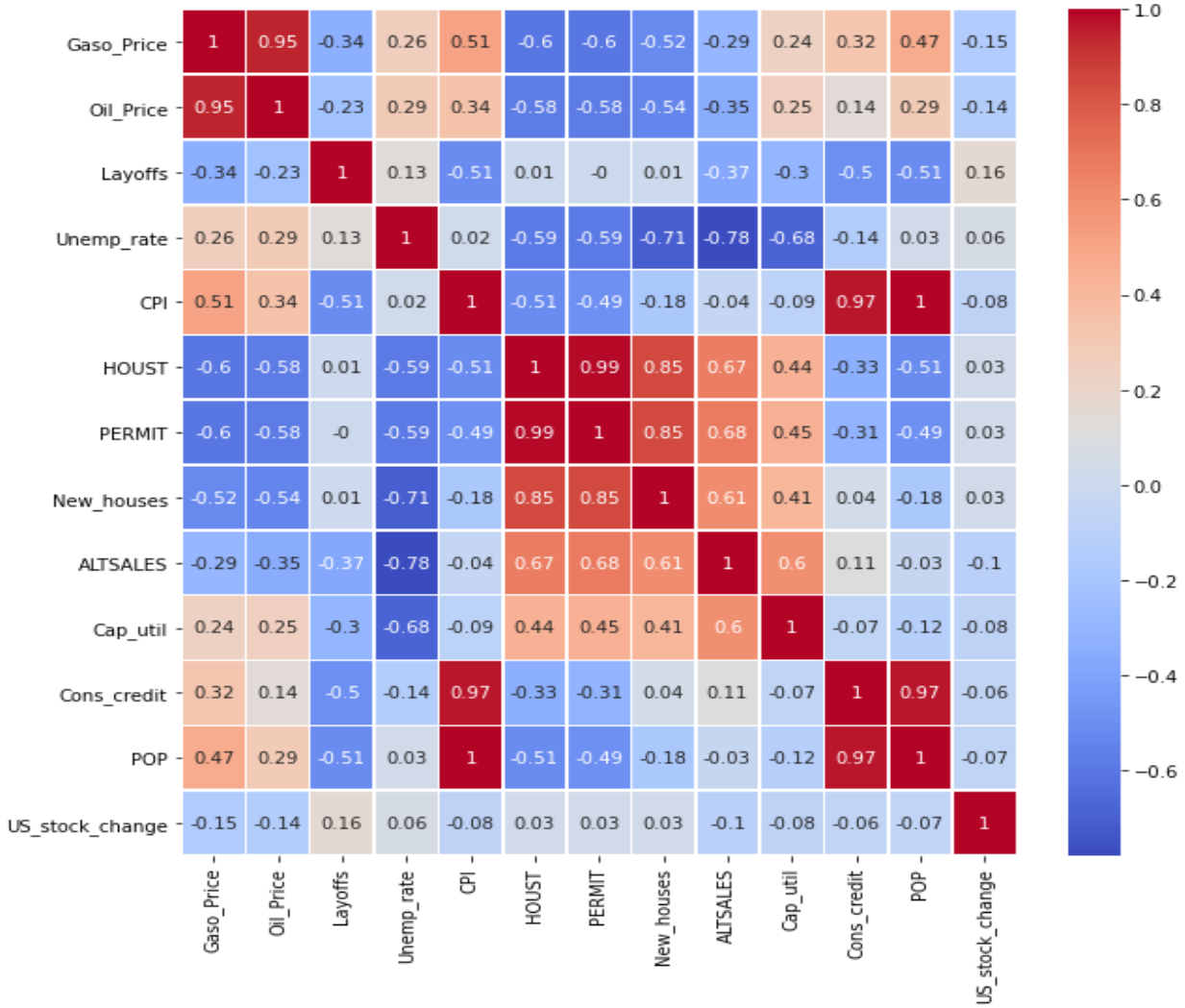

Figure 2 Correlation Matrix Heat Map

This study used multiple linear regression to predict crude prices and cluster analysis to identify patterns in the data. As far as building the multiple regression model, 70% percent of the data was randomly chosen to train the models and 30% for testing. A Scikit learn pipeline specifying the minmax feature scaler was used to build the multiple linear regression models. K-means and hierarchical agglomerative clustering methods were considered to identify data patterns. The analysis for both clustering methods assumed 7 clusters and used the minmax feature scaler. The cluster analysis with both methods excluded crude price and gasoline price which are considered dependent variables when building the clusters.

4. Results

The multiple linear regression analysis using the training data yielded a 0.91 R^2 so the model explained approximately 90% of the variability observed in crude oil prices. Using the same model to predict the test data set resulted in an R^2 of 0.86. Figures 3 provides charts of actual crude prices vs. predicted process for both training and test data sets highlighting the goodness of fit. Note that the test set (chart on the right) contained higher prices and those observations were the most difficult to predict based on the model trained model as the trained model set did not contain values exceeding 120 $/bbl.

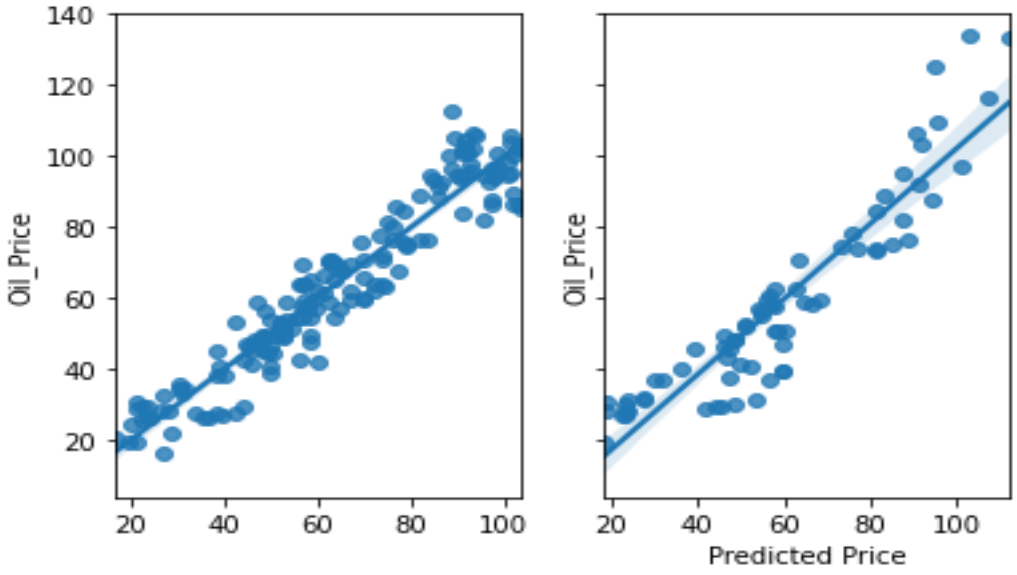

Figure 3 Regression Model Actual vs. Predicted Values

Table 2 below summarizes the average values for the different variables by cluster based on the K-means method. Focusing on clusters 2-4, which represent the high crude prices, it can be seen that the unemployment rate was on the high end of the spectrum for two of the clusters. In addition, variables representing new construction activity were on the low end of the spectrum for the same two clusters. It is important to point out that this analysis does not establish cause and effect so it is not known if high crude prices resulted in the relatively high unemployment rate or vice versa.

Clus_km	Gaso_Price	Oil_Price	Layoffs	Unemp_rate	CPI	HOUST	PERMIT	New_houses	ALTSALES	Cap_util	Cons_credit	POP	US_stock_change
0	2.5	52.8	1800.3	4.3	246.9	1232.2	1294.4	1073.8	17.1	76.9	3773.4	325355.6	-281.6
1	2.2	53.1	1936.9	5.2	194.3	1974.8	2042.7	1307.3	16.8	79.4	2244.8	295696.9	2026.4
2	3.4	91.7	1790.6	7.1	233.1	904.3	959.3	641.7	15.5	77.6	3036.1	316499.4	3089.9
3	3.0	86.4	1978.5	4.9	209.8	1234.3	1265.5	1101.1	15.5	80.3	2566.0	302578.6	461.6
4	2.8	77.2	2035.9	9.0	218.6	593.1	608.5	542.5	11.5	72.7	2612.5	309178.5	469.9
5	2.1	35.1	1739.3	10.4	258.3	1294.6	1369.7	1200.6	13.7	69.6	4143.0	330136.1	2104.0
6	1.4	27.5	2022.3	5.5	180.0	1701.2	1746.8	1029.9	16.8	75.6	1910.1	287758.9	2921.5

Table 2 K-Means Cluster Analysis

Table 3 also shows that the hierarchical clustering analysis also identified high prices associated with low new construction activity but the relationship with unemployment rate is not as clear.

cluster_	Gaso_Price	Oil_Price	Layoffs	Unemp_rate	CPI	HOUST	PERMIT	New_houses	ALTSALES	Cap_util	Cons_credit	POP	US_stock_change
0	2.0	44.6	1974.3	5.2	190.3	1771.8	1826.3	1172.3	16.7	78.1	2149.3	293353.9	2022.8
1	3.5	97.7	1834.7	6.5	229.3	965.2	1005.4	761.7	15.5	78.1	2968.3	313705.2	543.4
2	2.5	52.9	1800.6	4.4	246.2	1217.4	1282.1	1058.9	17.2	77.0	3743.3	325010.6	762.1
3	1.9	27.3	1926.7	13.0	256.3	1079.0	1180.0	1184.3	11.3	66.0	4129.1	329827.9	23509.3
4	2.4	64.7	2218.0	9.0	215.2	578.7	602.0	624.4	10.6	70.2	2587.6	307609.4	3593.8
5	2.2	37.8	1655.4	7.6	259.4	1418.8	1480.8	1213.6	14.7	72.6	4161.9	330209.7	-5514.2
6	3.4	91.2	1836.7	8.9	224.4	639.5	661.5	442.9	13.0	75.9	2693.3	311974.1	-1053.0

Table 3 Hierarchical Agglomerative Cluster Analysis

Figure 4 below shows house permits vs. unemployment rate for the observations in the data set with the clusters highlighted. The light blue cluster which is cluster 1 above shows relatively low new house permits and middle of the range for unemployment rate. Cluster 6, on the other hand, shows high unemployment rate and low new house permits.

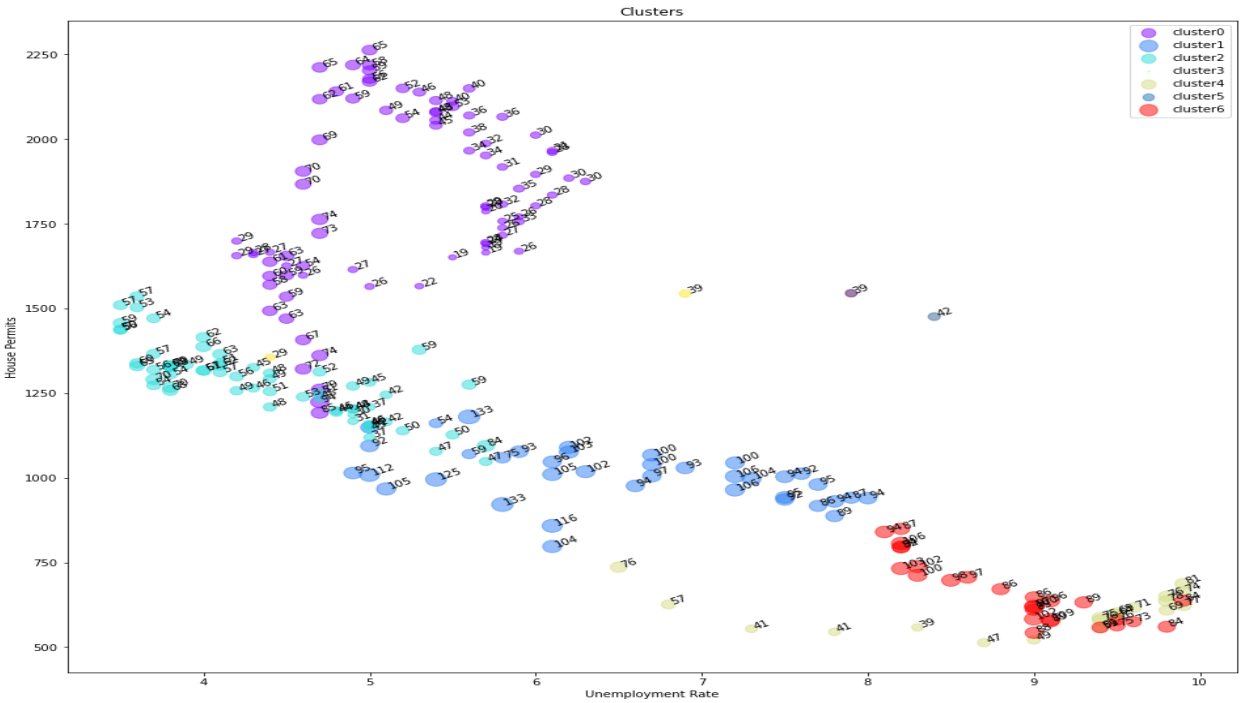

Figure 4 Hierarchical Cluster House Permits vs. Unemployment Rate by Cluster

5. Discussion

The analysis suggests a more direct relationship between gasoline prices and some of the economic variables when compared to crude price as evident from the correlation matrix. The analysis gave models with high predictive power but needs to be expanded to consider lagged terms to better identify leading indicators to generate predictions of future prices. The models built suggests high correlation with economic variables but causation relationship not established. As far as future work, adding more variables such as OPEC production, international petroleum stocks, and GDP for the US and international countries might further improve the model. The cluster analysis showed high prices associated with low new construction activity while the relationship with unemployment rate not as clear.

6. Conclusion

The analysis yielded a model with high predictive power but further work is required to identify the best leading indicators for use in forecasting future crude prices. Based on this analysis, it is recommended to conduct similar analysis for gasoline prices as the analysis suggests a more direct relationship between gasoline prices and some of the economic variables considered when compared to crude. The model built suggests a significant relationship between crude price and economic variables but causation relationship not established.

As far as future work, adding more variables such as OPEC production, international petroleum stocks, and GDP for the US and international countries might further improve the models. Additional work is also recommended to experiment with different number of clusters for both clustering methods. Other techniques such as CART regression trees should be considered to better account for non-linearities in the data. The model could also incorporate economic variables such as interest rates and a stock market index such as the S&P 500.