Forecasting Crude Oil and Gasoline 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 to process at their refineries and often engage in crude hedging programs 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 would also be interested in planning their future transportation logistics expansions.

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. 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 frequency of the data is monthly and it covers the years 2001 through 2020. The data will be used to build several models to predict crude oil and gasoline prices as a function of several economic variables.

FUTURE SUBMISSION

WORK IN PROGRESS

3. Methodology

- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

First step in the explanatory data analysis consisted of examining the values of both the dependent and the independent variables using the pandas describe method. Table 1 provides the summary generated. It shows a significant outlier for the layoffs column. Values of layoffs exceeding 5,000 were replaced with the median value of xx. Table 2 is the summary after the outliers were replaced.

A correlation matrix was also generated and provided in Table 3. Figure 1 shows the heatmap generated using the correlation matrix. As expected, gasoline prices were found to have the highest correlation with crude price.

Seventy percent of the data was randomly chosen to train the models and 30% for validation.

4. Results

- Results section where you discuss the results.

5. Discussion

- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

6. Conclusion

- Conclusion section where you conclude the report.