Predictive Analysis of Brent Crude Oil Prices Using MLR Models

Executive Summary

Oil markets are influenced by many different complex variables, which makes accurate forecasting difficult but valuable to many firms and organizations. This study is particularly relevant to energy and financial firms and policymakers. The study creates a multiple linear regression (MLR) model to predict the daily close price of Brent crude oil based on related commodity close prices of crude oil, natural gas, and heating oil. The hypothesis derived from the objective of this study is summarized as follows: The selected variables align effectively with the given data, and the MLR model demonstrates the potential to generalize and achieve high accuracy in predicting unseen data, supporting decision-making and risk management in energy markets.

The dataset used (Brent Crude Oil Dataset, 2024) includes 21 variables and 5,558 observations, which was large enough to allow for the data to be split into a training set (80%) and a test set (20%). The independent variables consisted of continuous and quantitative types related to daily commodity closing prices of Brent crude oil (target variable), crude oil, heating oil, and natural gas (predictor variables). The dataset was sourced from a GitHub repository that is publicly available, which combined data from the U.S Energy Information Administration and the Federal Reserve Bank of St. Louis (FRED). The analysis utilized several Python libraries. The most utilized include Pandas and NumPy, which were the tools used for data cleaning, which included outlier detection and dropping of outliers, addressing multicollinearity by dropping variables with high Variance Inflation Factor (IVF), which led to the dropping of Close Crude Oil variable. The retained variables were then standardized and used with Brent Crude Oil to create a multiple linear regression (MLR) model. The model produced an R-squared and Adjusted R-squared value of 97.4%, which indicates a strong linear relationship.

The model was then evaluated to check for heteroscedasticity and non-normal residuals using the Breusch-Pagan and Shapiro-Wilke tests. In the original model, the Breusch-Pagan test resulted in 122.84 statistics indicating heteroscedasticity in the residuals. The Shapiro-Wilke test yielded a 0.89 statistic, indicating that residuals deviated from normality. A log transformation was applied to the dependent variable Brent Crude Oil to address these issues. The new model was then tested using the Breusch-Pagan and Shapiro-Wilke tests, resulting in 182.688 and 0.932, respectively. The Breusch-Pagan score got worse, and the Shapiro-Wilke test got better but by a very small amount. It was decided to keep the original model; its results do not use the log-transformed model.

While this study gave us insight into the relationship between the dependent and independent variables, the study had limitations. One of the limitations included multicollinearity with the independent variables that led to the removal of Close Crude Oil, which could have excluded a helpful variable that limited the model’s explanatory power. The Breusch-Pagan test revealed heteroscedasticity even after a log transformation, violating a key MLR assumption. The Shapiro-Wilk test showed that the model showed residual non-normality and could not resolved by a log transformation of the model. While outlier removal helped improve homoscedasticity and residual normality, it excluded variable data points. Data points that could have tied back to geopolitical events that the model may not account for now.

With these limitations in mind, suggestions for future study on this topic and dataset would be to add a macroeconomic or geopolitical variable such as GDP or inflation. This would help capture external factors that influence commodity prices. Including these variables could improve the model, specifically during geopolitical events. Another suggestion for further research is implementing Weighted Least Squares (WLS) in the model. This could help with heteroscedasticity and improve the model's reliability.

The study provides significant and measurable benefits, particularly for organizations in the energy and financial sectors. By achieving a high R-squared value of 97.4%, the Multiple Linear Regression (MLR) model offers a useful tool for predicting the daily closing price of Brent Crude Oil based on the closing prices of related commodities like Heating Oil and Natural Gas. This predictive capability enables energy firms, traders, and oil analysts to make decisions on production, inventory management, and trading strategy. For example, improving the prediction accuracy of the price of Brent Crude Oil by even 1% could lead to a significant increase in profits, especially when firms operate on large trading volumes.

-No sources used.