Predicting Refined Oil Prices: A Data Mining Approach Based on Crude Oil Prices

Xinyi Liu (V00850820)

Jingyao Yu (V00972776)

Ran Ding (V00970789)

Abstract

This research focuses on the application of data mining techniques to predict refined oil prices using crude oil prices. Understanding the fluctuation in refined oil prices is a complex problem due to various factors influencing it, among which the price of crude oil plays a significant role. In this study, historical data of crude and refined oil prices were gathered and processed using advanced data preprocessing techniques. Subsequently, various data mining algorithms were employed to establish predictive models. The models were evaluated based on their forecasting accuracy. The findings reveal that there is a substantial correlation between crude oil and refined oil prices, and the proposed data mining models demonstrated a high level of accuracy in forecasting refined oil prices. This research can provide valuable insights for stakeholders in the oil industry, contributing to more effective strategic planning and decision making. Future work may include incorporating additional factors influencing oil prices and enhancing prediction models for better accuracy.

Table of Contents

1. Introduction

The petroleum industry, a cornerstone of the global economy, is characterized by its substantial price volatility. This variability introduces a high degree of uncertainty and risk for all stakeholders involved, including oil companies, investors, and even national governments. In this context, the capability to accurately predict prices of various petroleum products becomes a critical need, providing a foundation for more strategic decision-making processes and investment planning.

The current project aims to address this need by harnessing the power of machine learning techniques to forecast the prices of different petroleum products, specifically gasoline, diesel, heating oil, and jet fuel. The primary focus of our study is to examine the predictive power of historical West Texas Intermediate (WTI) crude oil prices. Crude oil prices, being a significant determinant of refined oil prices, serve as a crucial input to our forecasting models.

Alongside crude oil prices, other relevant factors that potentially influence the prices of these petroleum products will be taken into account. These factors might include geopolitical developments, changes in global oil demand and supply, fluctuations in exchange rates, and alterations in governmental policies among others. Acknowledging the multifaceted nature of petroleum pricing, we strive to incorporate these elements into our analysis to improve the robustness and reliability of our forecasting models.

To achieve our goal, we plan to construct and evaluate a series of regression models. These models will be trained and validated using historical data, with their performance gauged based on their ability to accurately forecast the future prices of the aforementioned petroleum products.

As this research unfolds, we anticipate encountering challenges relating to data quality, model selection, and potential overfitting. Nonetheless, we are committed to addressing these challenges head-on, learning from them, and ensuring our results are as robust, reliable, and useful as possible.

2. Data Preprocessing

In order to create accurate and reliable predictive models, we first embarked on a thorough data collection and preparation process. Our data was collated from multiple publicly accessible databases to ensure a robust and wide-ranging dataset that would enhance the efficacy of our models.

2.1 Raw Data Analysis

The data collected included variables such as West Texas Intermediate (WTI) crude oil prices, gasoline prices, diesel prices, heating oil prices, jet fuel prices, and the Consumer Price Index (CPI). This diverse selection of variables was chosen to ensure our model accounts for a wide range of factors that could potentially influence refined oil prices. These datasets spanned a comprehensive time frame of more than two decades, from January 1, 1998, to June 1, 2023. Here is an example of how we retrieve and process WTI data.

2.2 Data Transformation

Once the data was collected, it underwent a meticulous preparation phase to ensure it was in the optimal format for processing by our machine learning models. The following steps were undertaken to prepare the data:

1. **Date Field Conversion:** The date fields in our dataset were initially in a format unsuitable for our processing needs. We converted these fields into a date type format to ensure proper alignment and correlation with our other variables.
2. **Data Filtering and Cropping:** To focus our analysis on the most relevant data, we filtered and cropped our datasets according to the specified time range. This allowed us to remove any extraneous data points that fell outside of our study period, ensuring a cleaner, more focused dataset.

3. **Missing Values Imputation:** Real-world datasets often contain missing values, which can hinder the accuracy of predictive models. In our case, we chose to fill these missing values using the K-Nearest Neighbors (KNN) Imputation method. This method replaces missing values with the mean value of the k nearest neighbors, thereby maintaining the overall distribution and relationships within the dataset.
4. **Data Sorting and Reindexing:** Finally, we sorted the data chronologically and reindexed it to provide a streamlined structure, which would aid subsequent data manipulation and processing steps.

2.3 Data Visualization and Exploration

Understanding the nature of our data and identifying potential relationships between different variables is an essential precursor to building effective predictive models. To facilitate this understanding, we implemented a comprehensive data exploration and visualization process, harnessing the power of various data analysis libraries.

We primarily used the Seaborn library, a powerful Python library designed for data visualization, to generate a series of scatter plots. These scatter plots illustrated the relationships between West Texas Intermediate (WTI) crude oil prices and the prices of different petroleum products, namely gasoline, diesel, heating oil, and jet fuel. The pairplot function of the Seaborn library was particularly instrumental in this task. It allowed us to visualize pairwise relationships across the entire dataset, making it easier to discern patterns and correlations at a glance.

To quantify the relationships observed in the scatterplot, we calculated the correlation coefficients between crude oil prices and the prices of various petroleum products. Figure 1 shows an example of the correlation between crude oil prices and New York gasoline prices.

Our analysis revealed a strong positive correlation between WTI crude oil prices and the prices of gasoline, diesel, heating oil, and jet fuel. This suggests that as crude oil prices increase, we can generally expect the prices of these petroleum products to increase as well, and vice versa.

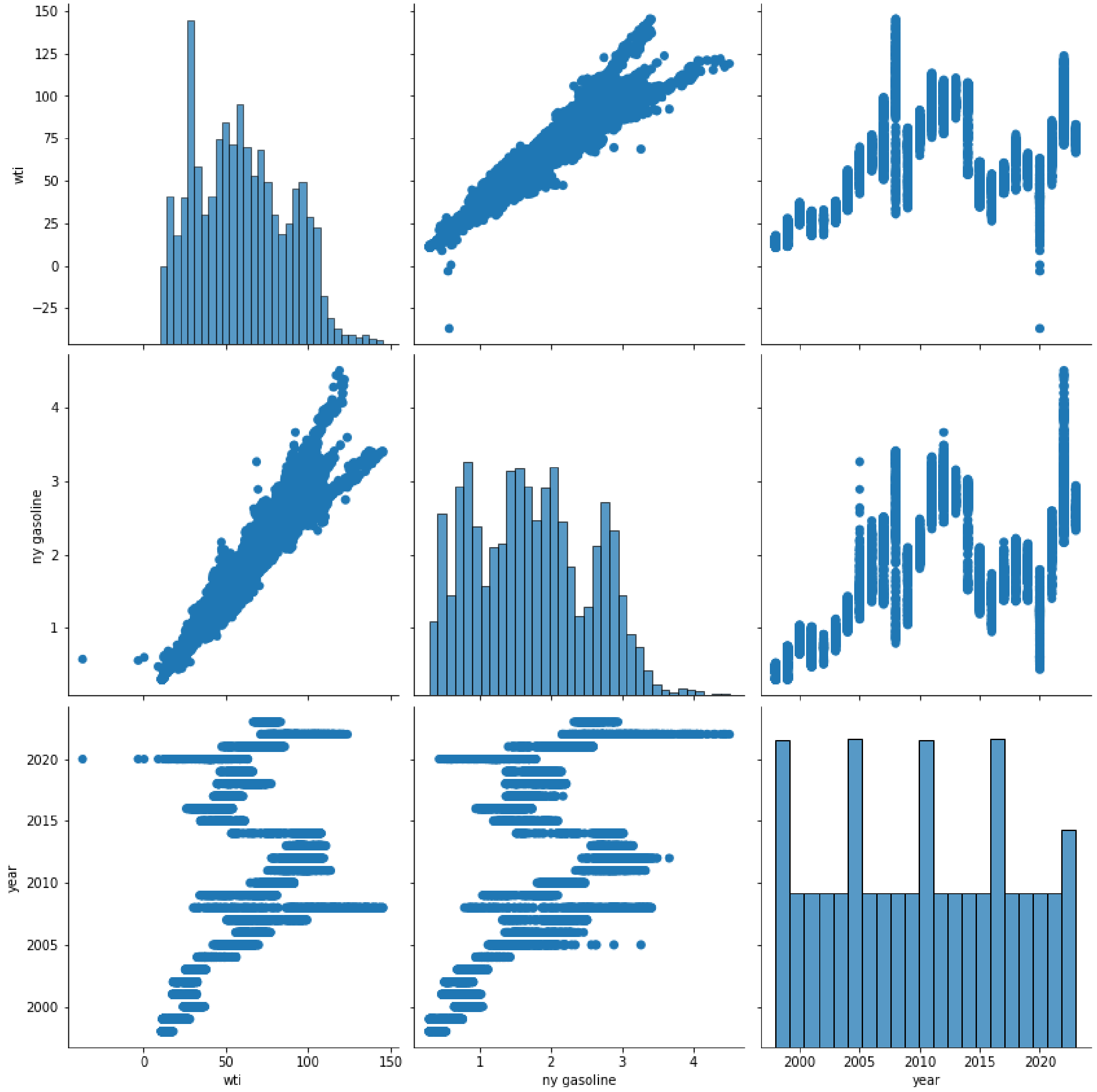

Figure 1. The plot gram for WTI price and New York gasoline price

2.4 Splitting Data for Training

Finally, we divided our data into separate training and test datasets. The training dataset, which included the majority of our data, was used to train our machine learning models. The test dataset was set aside and used to evaluate the performance of our models. This approach ensures that our evaluation is robust and simulates how the models would perform when deployed in the real world with unseen data.

3. Data Mining

To forecast the prices of various petroleum products, a sophisticated data mining process was employed. Our dataset was divided into a training set and a test set, a common practice in machine learning to ensure the robustness and generalizability of our models.

In order to harness the strengths of various machine learning algorithms and make our predictions more reliable, we employed an ensemble of regression models. These models included:

1. **GradientBoostingRegressor:** This model uses the gradient boosting framework, which constructs new predictors that aim to correct the residual errors of the preceding predictors, hence optimizing the predictive accuracy.
2. **RandomForestRegressor:** Based on the random forest algorithm, this model aggregates the predictions of numerous individual decision trees, which helps reduce overfitting and enhances model stability and accuracy.
3. **LinearRegression:** As one of the simplest and most widely used predictive modeling techniques, linear regression models the relationship between a dependent variable and one or more independent variables.
4. **DecisionTreeRegressor:** This model uses a decision tree approach, where the value of the target variable is estimated by learning simple decision rules from the data features.

Each of these models was trained on the training set, and then used to make predictions on the test set. This provided a comparative analysis of their performance and helped us identify the best model for our requirements.

To assess the performance of each model, we calculated the following evaluation metrics:

1. **Mean Absolute Error (MAE):** This represents the average of the absolute differences between the predicted and actual values. It provides a straightforward measure of prediction error rates.

2. **Root Mean Square Error (RMSE):** RMSE is the square root of the average of the square differences between the predicted and actual values. It is particularly useful when large errors are undesirable.
3. **R2 Score (Coefficient of Determination):** The R2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well future samples are likely to be predicted by the model.

By comparing these evaluation metrics for each of our models, we were able to identify the most accurate and reliable approach for predicting the prices of different petroleum products.

4. Results and Discussion

Our data mining process yielded distinct results for the prediction of different petroleum product prices, with different regression models demonstrating superior performance for different types of petroleum products.

In the case of gasoline price forecasting, the Linear Regression model delivered the most accurate results. It yielded the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), indicating fewer errors in price prediction compared to other models. Additionally, this model achieved the highest R2 score, signifying that it could account for a substantial proportion of the variance in gasoline prices.

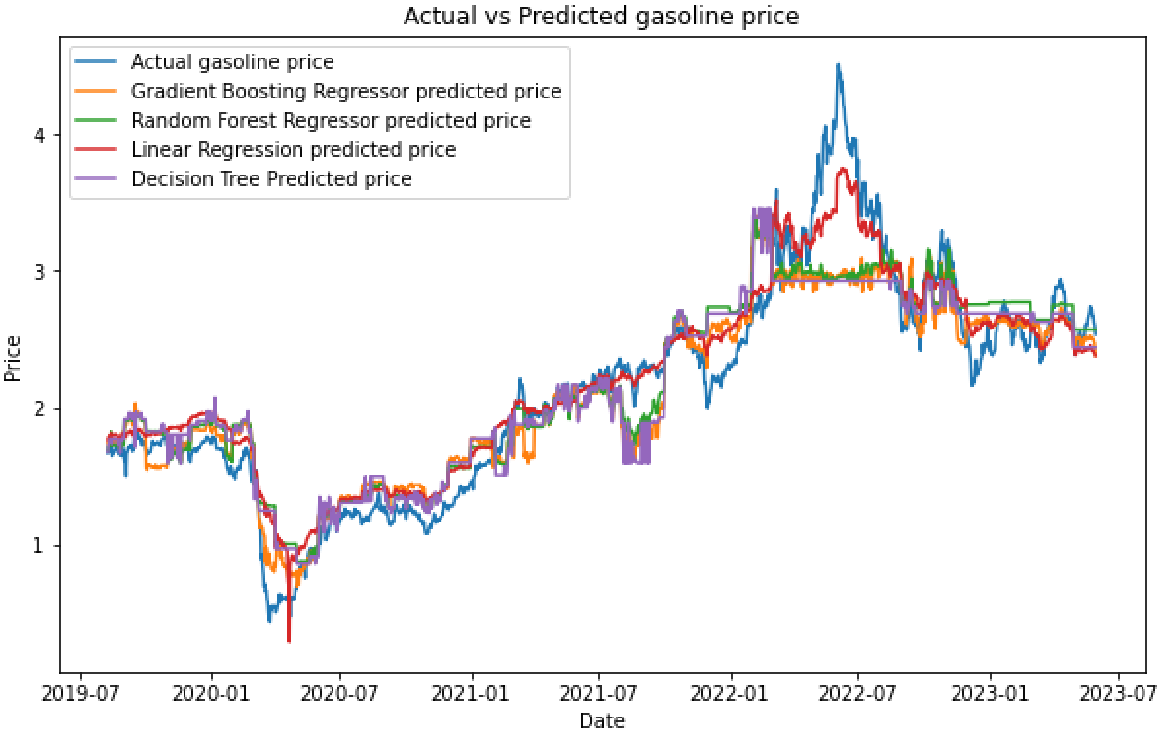

Figure 2. The comparison of different models on gasoline price prediction

Petroleum	**Model**	**MAE**	**RMSE**	**R2 score**
gasoline	GradientBoostingRegressor	0.219918	0.313273	0.844162
gasoline	RandomForestRegressor	0.228270	0.324499	0.832794
gasoline	LinearRegression	0.156131	0.200802	0.935973
gasoline	DecisionTreeRegressor	0.244917	0.342985	0.813200

Table 1. The index of different models on gasoline price prediction

When forecasting diesel prices, still the Linear Regression emerged as the superior model. It showed the lowest MAE and RMSE, highlighting its accuracy and reliability. Its R2 score was also highest amongst the tested models, indicating its robust predictive power for diesel prices.

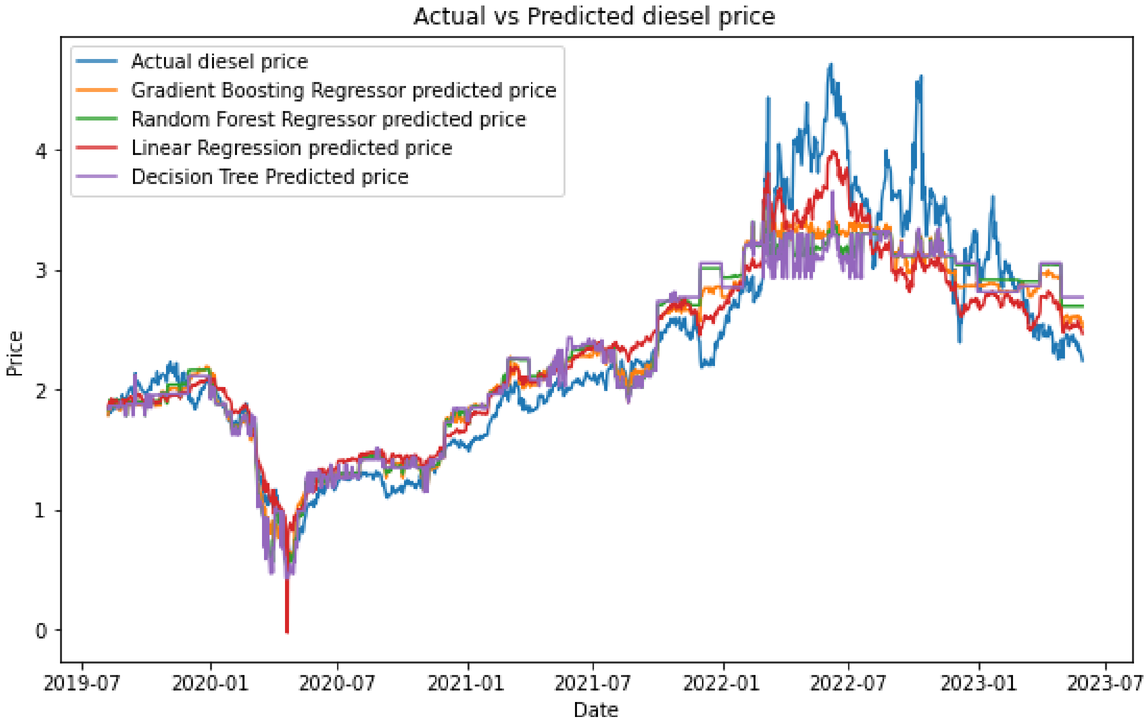

Figure 3. The comparison of different models on diesel price prediction

Petroleum	Model	MAE	RMSE	R2 score
diesel	GradientBoostingRegressor	0.244894	0.336128	0.863324
diesel	RandomForestRegressor	0.278596	0.380213	0.825121
diesel	LinearRegression	0.225813	0.293983	0.895449
diesel	DecisionTreeRegressor	0.293786	0.403105	0.803429

Table 2. The index of different models on diesel price prediction

For heating oil price prediction, the Linear Regression model once again demonstrated superior performance. It exhibited the lowest MAE and RMSE and the highest R2 score, confirming its effectiveness in predicting heating oil prices.

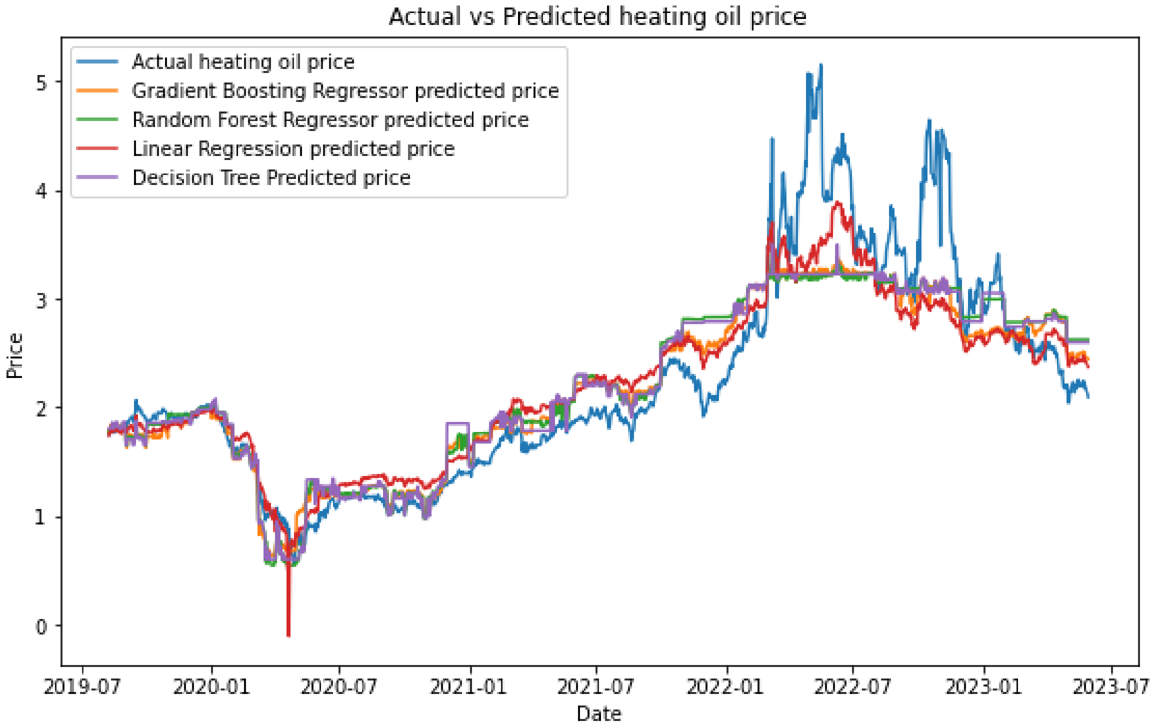

Figure 4. The comparison of different models on heating oil price prediction

Petroleum	**Model**	**MAE**	**RMSE**	**R2 score**
heating oil	GradientBoostingRegressor	0.292133	0.423356	0.813074
heating oil	RandomForestRegressor	0.303230	0.443923	0.794471
heating oil	LinearRegression	0.278477	0.397574	0.835148
heating oil	DecisionTreeRegressor	0.299173	0.439083	0.798928

Table 3. The index of different models on heating oil price prediction

In the realm of jet fuel price forecasting, the Linear Regression still proved to be the most effective. It achieved the lowest MAE and RMSE and the highest R2 score, indicating strong performance in predicting jet fuel prices.

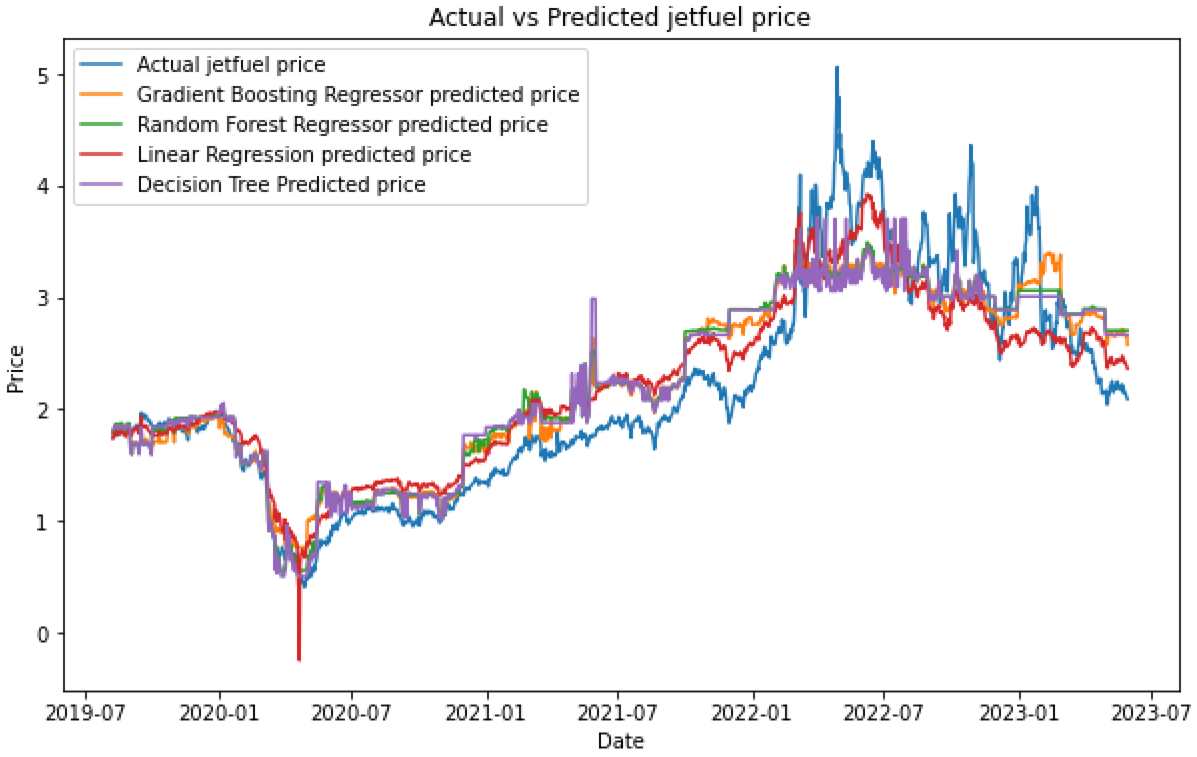

Figure 5. The comparison of different models on jet fuel price prediction

Petroleum	Model	MAE	RMSE	R2 score
jetfuel	GradientBoostingRegressor	0.300307	0.377546	0.840488
jetfuel	RandomForestRegressor	0.308715	0.391514	0.828467
jetfuel	LinearRegression	0.282744	0.347626	0.864768
jetfuel	DecisionTreeRegressor	0.311898	0.402492	0.818712

Table 4. The index of different models on jet fuel price prediction

Given these results, we recommend adopting the Linear Regression model for forecasting all kinds of refined oil (gasoline, diesel, heating oil and jet fuel). By utilizing these models, companies and investors can improve the accuracy of their price forecasts, thereby enhancing their decision-making process and potentially increasing profitability in an unpredictable market.

In summary, this study demonstrates the effective application of machine learning models in the prediction of petroleum product prices. Further improvements could be achieved by tuning model parameters, exploring additional features, and applying more advanced machine learning techniques. Nonetheless, the models developed herein provide a robust foundation for future studies and practical applications.

5. Conclusion

This study has effectively demonstrated the viability of using machine learning models to forecast the prices of various petroleum products. By leveraging historical data and applying sophisticated regression algorithms, we have successfully built and evaluated predictive models tailored to each type of petroleum product. These models not only provide insight into past price trends but also equip us with a reliable tool to forecast future price movements.

While our models have shown promising results, it is imperative to acknowledge that price prediction remains a challenging endeavor, shaped by a myriad of interlinked factors. The complexity and volatility of the global petroleum market, influenced by geopolitical, economic, and environmental events, present significant hurdles for any predictive model.

Therefore, our recommendation for future work involves continued refinement and optimization of these models in real-world applications. This may include incorporating more granular data, considering additional influencing factors, and integrating more advanced or specialized predictive algorithms. Furthermore, we encourage the exploration of other features that may have been overlooked in this study but could potentially enhance the predictive power of the models.

In conclusion, while the models we have developed are robust and reliable within the scope of this project, their continued evolution will be key in maintaining their relevance and accuracy in the ever-changing landscape of the petroleum industry.