**Stock Analysis and Prediction Team**

**members:**

Anusha Kurra- 11578831

Varsha Jampala- 11693177

Neeraj Reddy Aluka- 11694856

Sree Kohali Patel- 11694857

Rashmitha Chikoti- 11706417

# Abstract

Big data analytics play a crucial role across various sectors by accurately predicting and analyzing extensive datasets, revealing otherwise concealed valuable information. A proposed method involves a resilient data pipeline built on Spark, capable of analyzing data of any magnitude and nature. This approach focuses on assessing four chosen US stocks, leveraging the Apache Spark framework to manage large datasets through distributed storage and processing. The stocks from the US market undergo analysis by dividing their daily gain information into training and test datasets, enabling the prediction of stocks with substantial daily gains using Spark's Machine Learning module.

**Introduction:**

In order to create a more accurate model for stock price prediction, the objective of this study is to investigate the correlation between oil prices and the stock prices of significant oil businesses.

This code illustrates how to predict the stock price of Capricorn Energy PLC, PetroChina Company Limited (PCCYF), Equinor ASA (EQNR.OL), BP p.l.c. based on oil prices using machine learning techniques. The CSV with the data on oil prices is loaded first, then each stock price file.

Subsequently, the information is converted and combined into a single master dataframe. We employed a number of charting techniques, including spark, numpy, matplotlib, random forest regressor, linear regression, pandas to show the data. Lastly, models that forecast the prices of Capricorn Energy PLC, PetroChina Company Limited (PCCYF), Equinor ASA (EQNR.OL), BP p.l.c. shares are created using supervised learning techniques like random forest and linear regression. Every parameter's significance is evaluated as well.

**Description of model approach:**

Since there are no null values or outliers, we did not employ any data processing procedures. Thus, we compared the effectiveness of two distinct machine learning models in predicting the share price using Python and a variety of libraries, including Pandas, Numpy, and Scikit-Learn. We then used Matblotlib libraries to visualize the data for additional data analytics and used scatterplot graph to plot the values of trained data and predicted values for linear regression and random forest models. First, we will apply linear regression to train and test the data.

After that, Random Forest Regression will be used to train and test the data. The share price of four firms in the same sector (BP.L","CNE.L","PCCYF","EQNR.OL). The mean squared error (MSE), as presented in the IEEE publication that is cited, will be used to evaluate each model's performance.

# Problem Statement and existing solution

**Problem Statement:** Develop a system leveraging big data analytics for tracking and predicting daily gains in US oil stocks. For that we selected four stock datasets in the US oil fund from yahoo finance and energy information administration. Implement a machine learning module in Spark to analyse and predict high daily gains, while ensuring effective division of daily gain data into training and test datasets. Detect and prevent illegal activities such as money laundering and financial fraud in the process.

**Existing Solution:** Implementing a Spark-based data pipeline provides a robust solution for scalable and efficient analysis of US oil stocks. Leveraging big data analytics and network analytics, the system utilizes Spark's machine learning module for high-frequency trading analytics. This comprehensive approach ensures the detection and prevention of illegal trading activities, contributing to a secure and insightful financial market environment.

# Comparative analysis

we wished to develop a machine learning model to predict the future crude oil prices using the USO(The United

State Oil Fund,USO) data and understand which set of features are better in prediction using Spark framework.

Some basic knowledge of the USO and Spark framework are reviewed in this section. In this study, we employ various regression algorithms to identify the most suitable model for predicting video views. The algorithms under consideration are:

Linear Regression: Because of its ease of use and interpretability, linear regression is a parametric model that presupposes a linear relationship between independent and dependent variables. But because it depends on a linear connection and is susceptible to outliers, it might not adequately represent complicated data structures.

Random Forest: Conversely, the RandomForest Regressor is an ensemble learning method that creates a large number of decision trees while it is being trained. It manages complex data relationships with ease and excels in prediction accuracy. By avoiding overfitting, this ensemble method ensures reliable predictions. However, because it combines several trees, it loses interpretability, and the training procedure can be computationally taxing for large datasets.

Apache Spark: Apache Spark is a powerful open-source distributed computing system that enables high-speed data processing and analytics across large-scale datasets. It offers flexibility, speed, and a versatile set of tools for various data processing tasks, utilizing in-memory computation to enhance performance significantly.

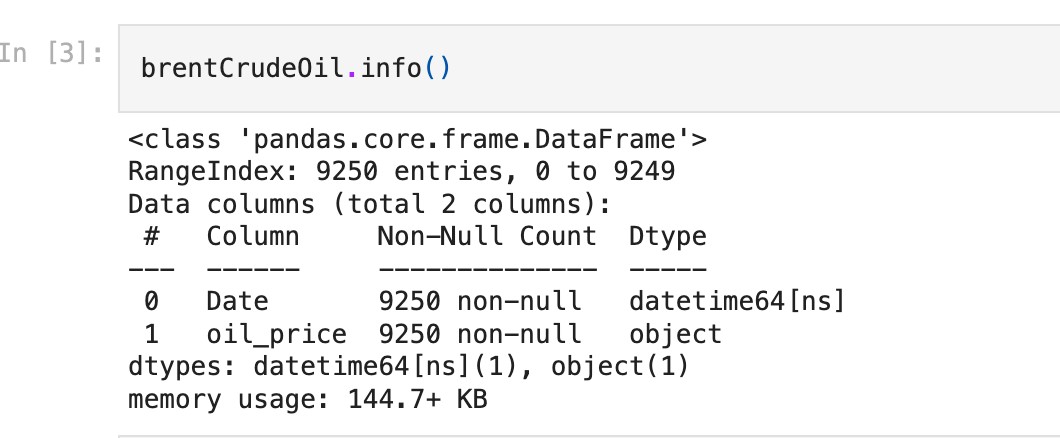
# Input dataset

After examining the references, we were unable to obtain the dataset that the author had utilized. The

U.S. Energy Information Administration provided it to us. The daily date and oil price are included in an CSV file that makes up the main dataset for oil. An CSV file with the stock data is supplied. We will load, read, and then convert the data before creating the master data frame.

Two sources of data were used in this investigation. The daily price of Brent crude oil from May 1987 to November 30, 2023 is the first. Four significant oil companies' stock prices make up the second part of the dataset Capricorn Energy PLC, PetroChina Company Limited (PCCYF), Equinor ASA (EQNR.OL), BP p.l.c. are the companies. All the data set contains values of 30,000+.

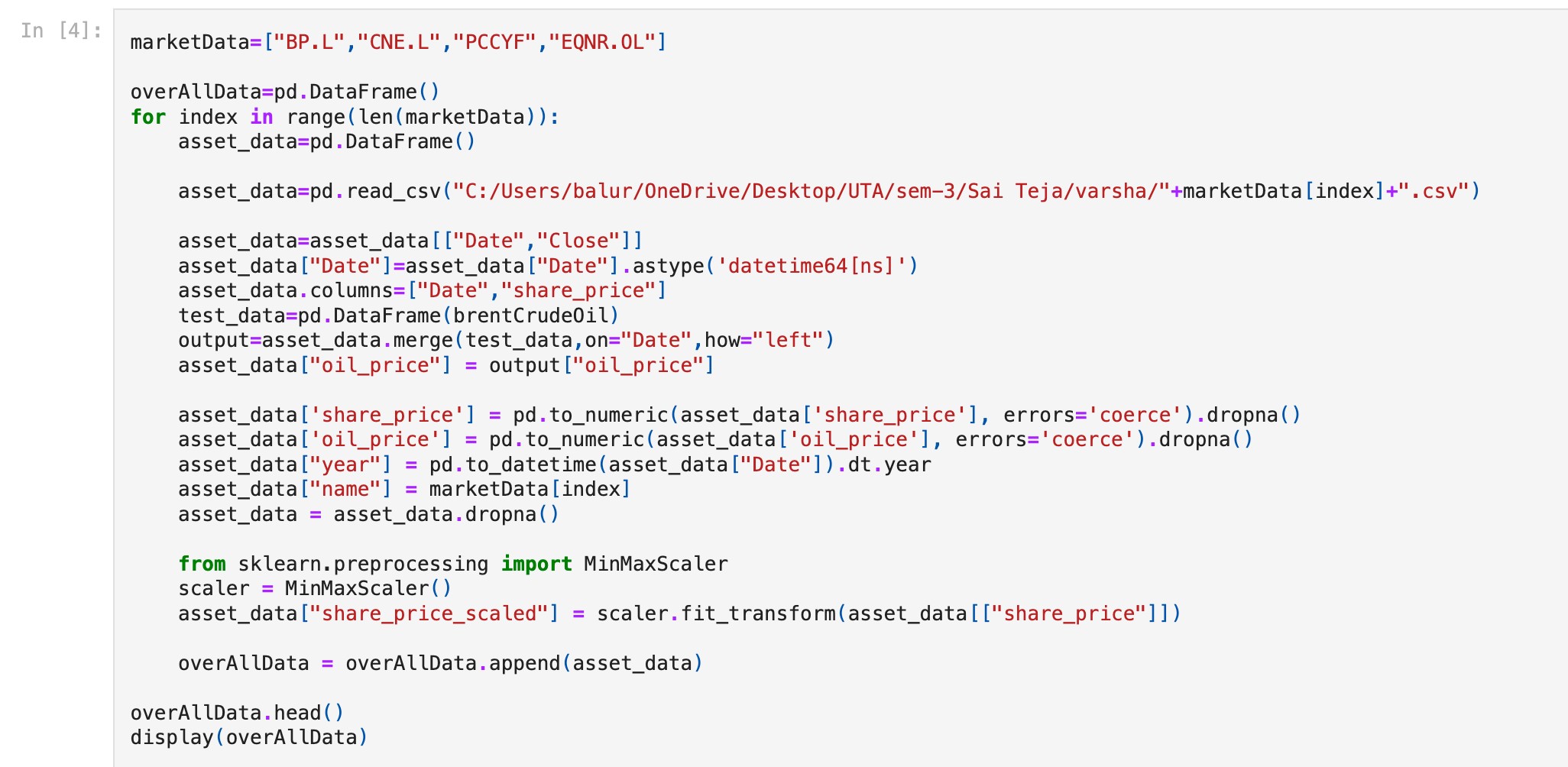




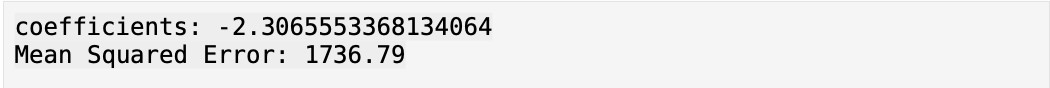
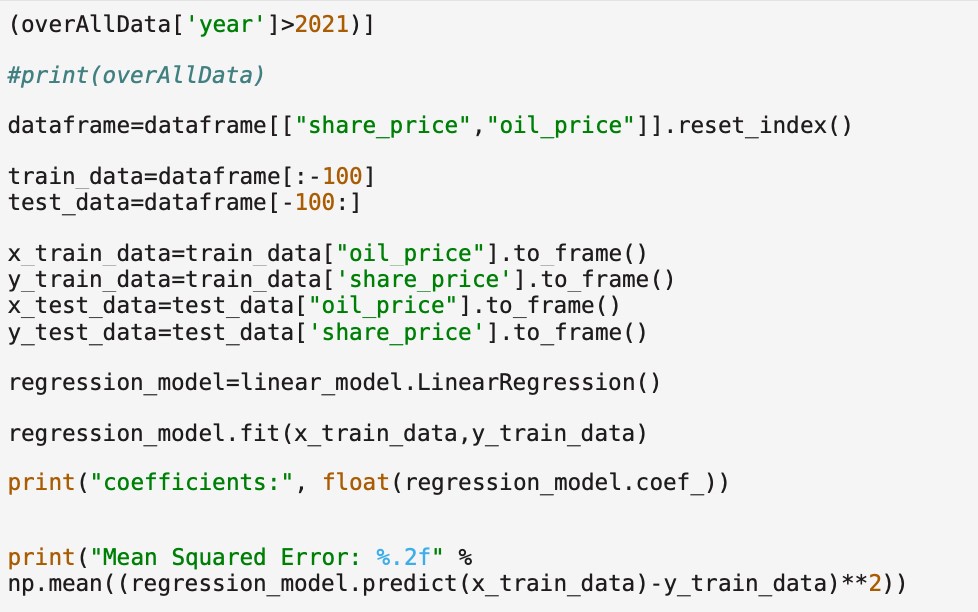
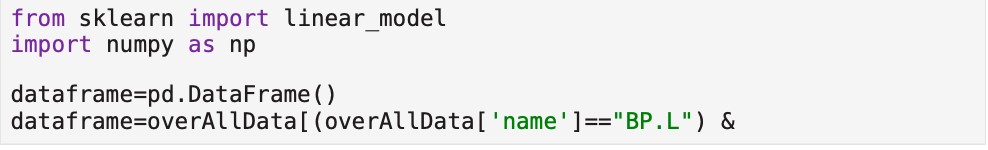
**Code screenshots:**

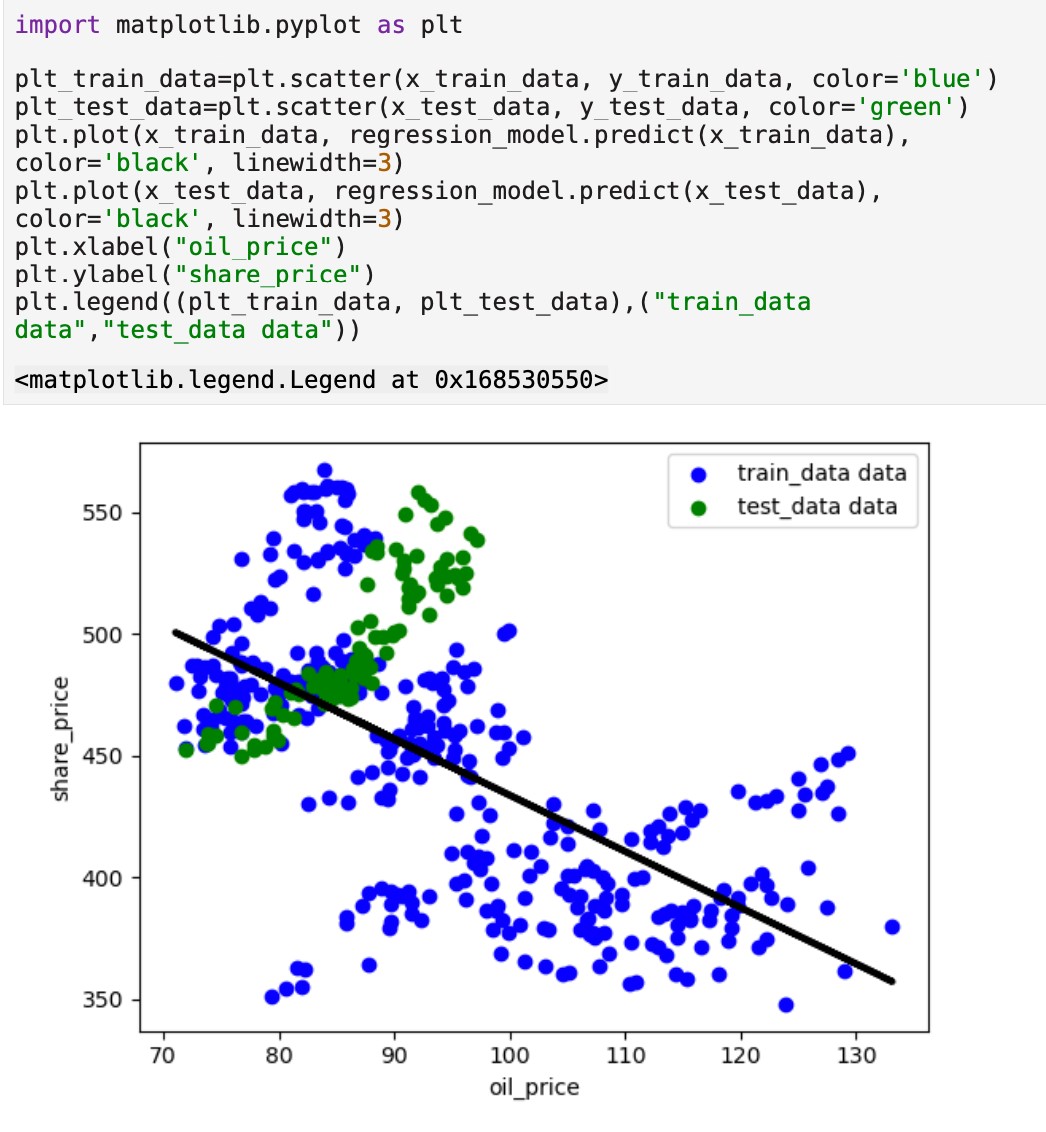
We have developed our project in jupyter notebook.

Now we have just loaded, transformed and checked our oil data. Reading and transforming the share price data before starting analysis.



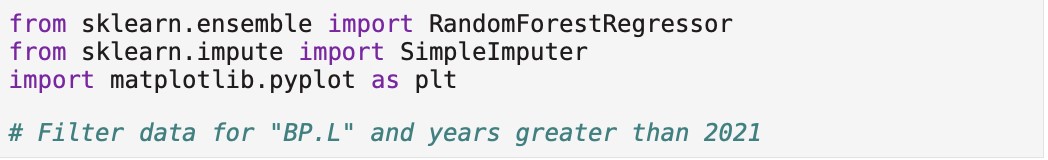
We will then use the supervised learning to build a simple linear regression model. The goal is to assess prediction of data from the last 100 days using data trained in 2016/17. The train data is the data used to build the model, and the test data is the data we are attempting to predict.





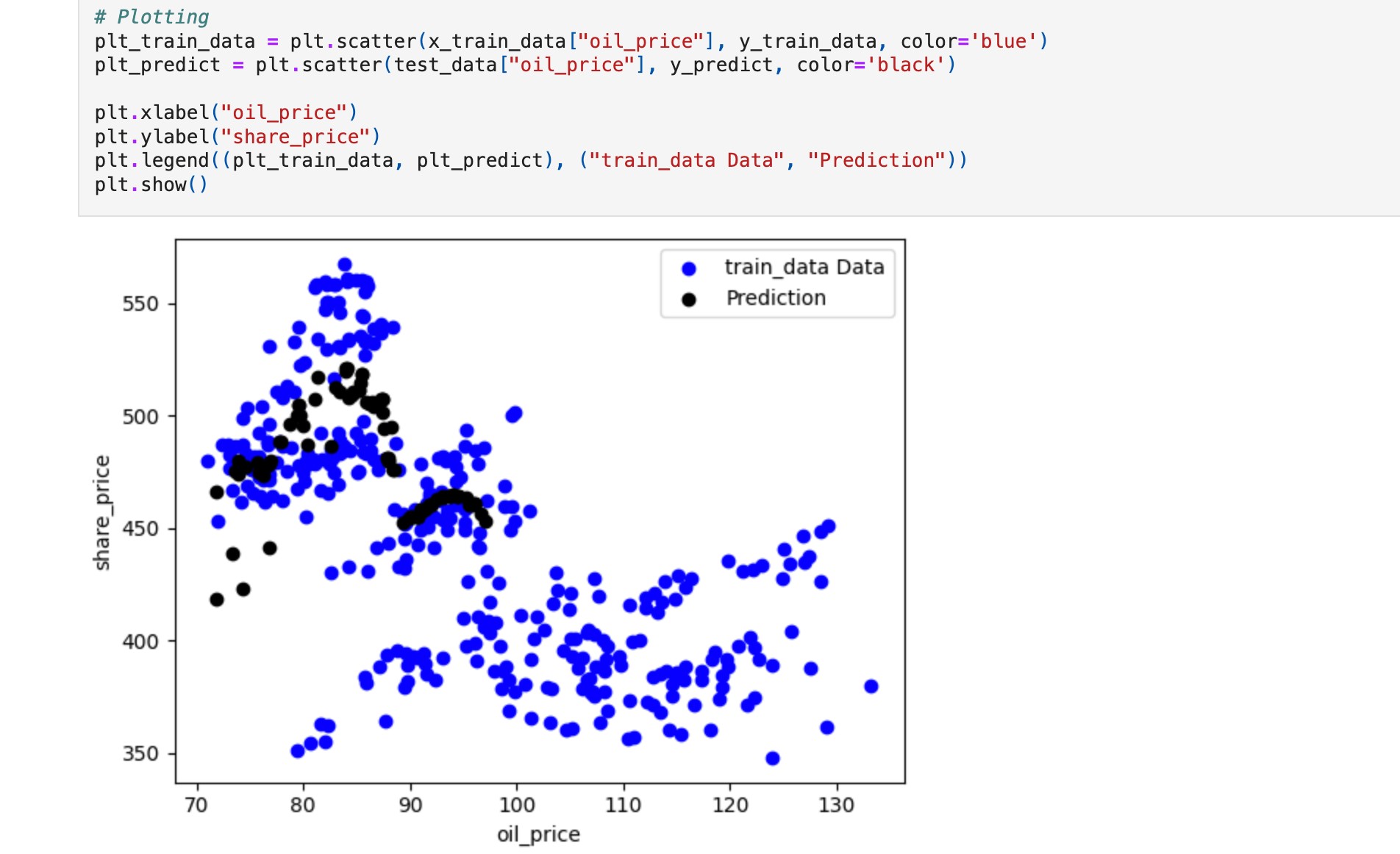
The approximation of how linear regression is fitted and attempts to predict a result from test data is displayed in the graph above. The statistics suggesting increasing oil prices seems to be far wrong. The mean square for this forecasting technique is 74.49. Let's examine the results of a more advanced approach on the matter. We'll employ the random forest in our job.

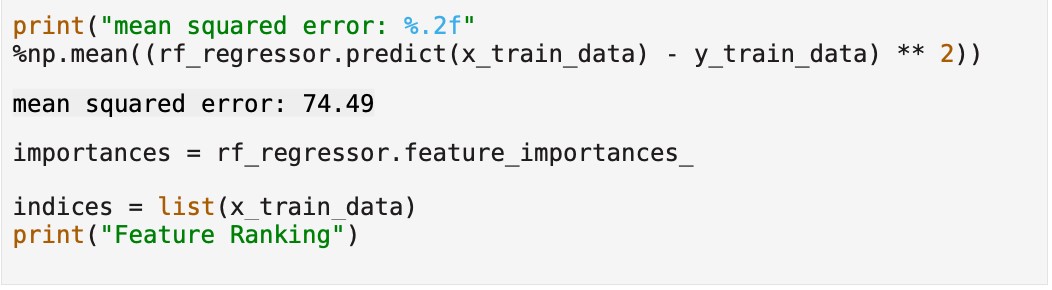
## Random Forest on share price vs oil price





Here we are going to look at how well the data has fitted.



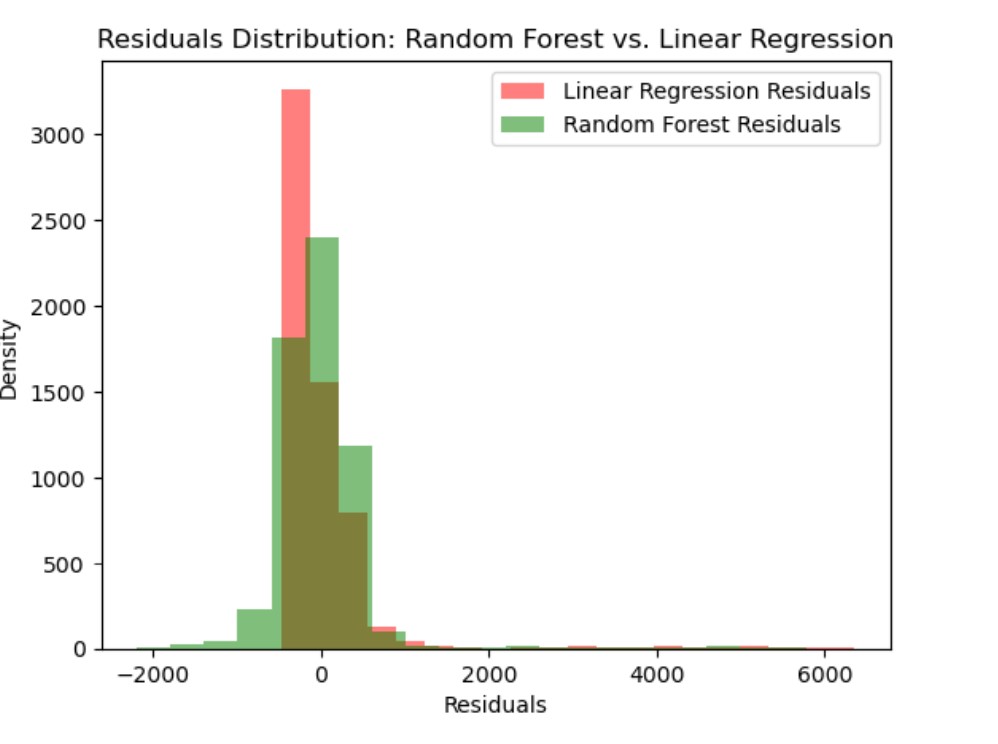


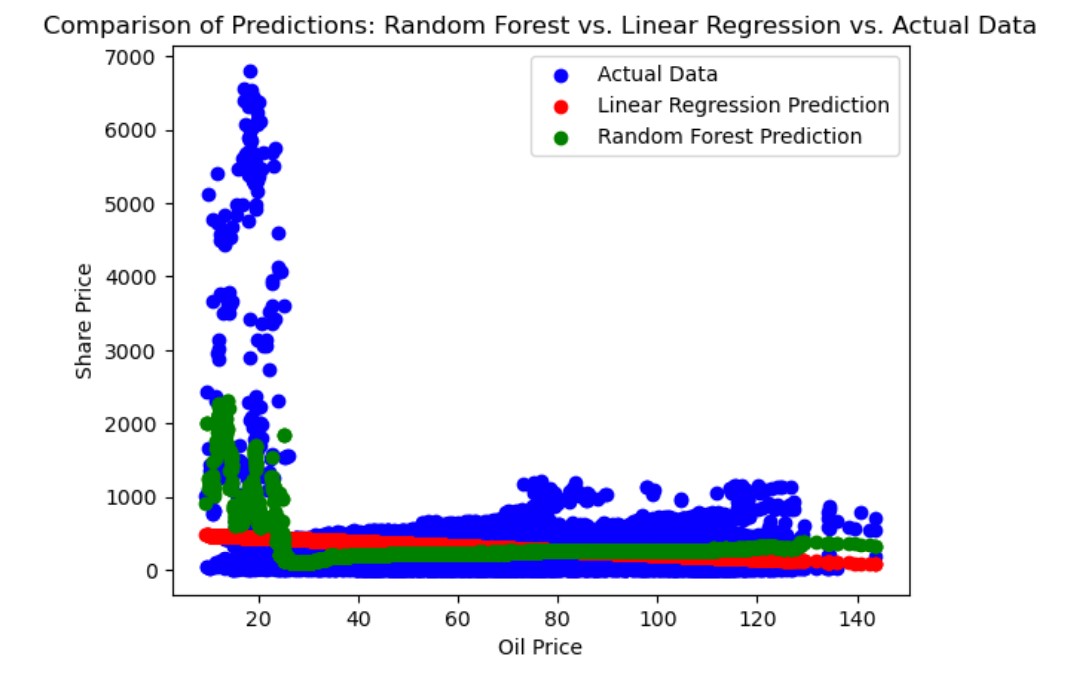
This improves the forecast based on the test data significantly. It is evident from the mean squared error that we were able to reduce the error from 1736.79 to 74.49. Compared to linear regression, this is twenty Emes less.

Results:

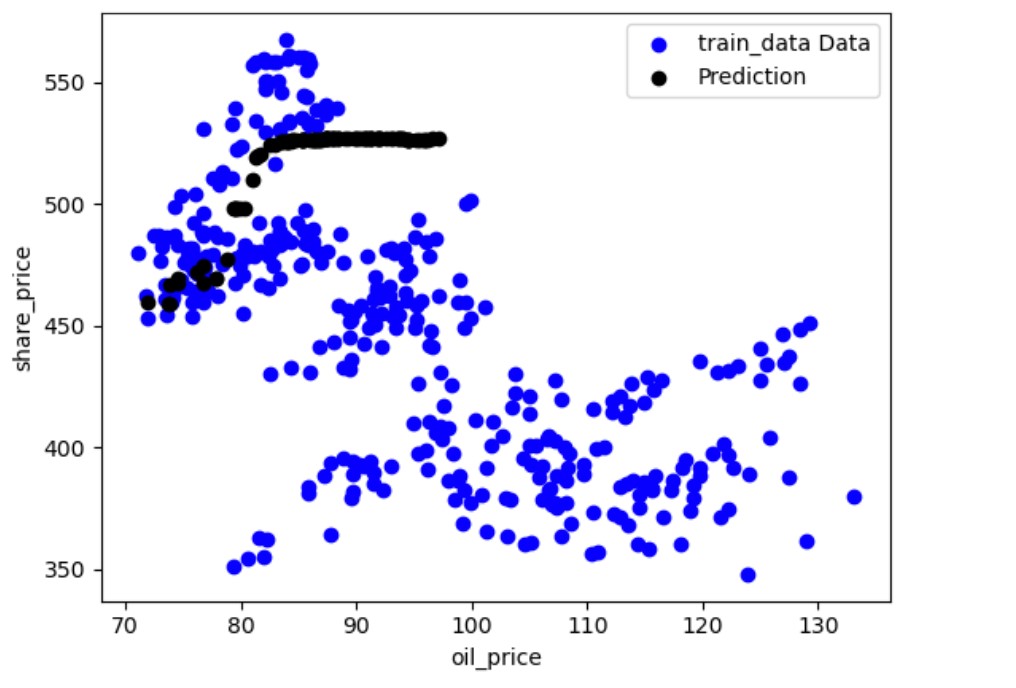
According to the distribution, oil prices over the last five years have fluctuated between

$30 and $60 per barrel and $100 and $120 per barrel. Not a lot of information between $60 and $100 a barrel... For 2016 and 2017, we also see a very favorable link between share price and oil price. Take note that 2014 was the watershed year for this firm; we will find out whether this is typical of other organizations in the future. There has been a shift in the correlation patterns and increased data variability in 2014. That line marked the transition between two distinct market behaviors and a new regulatory environment inside the organization as a whole.

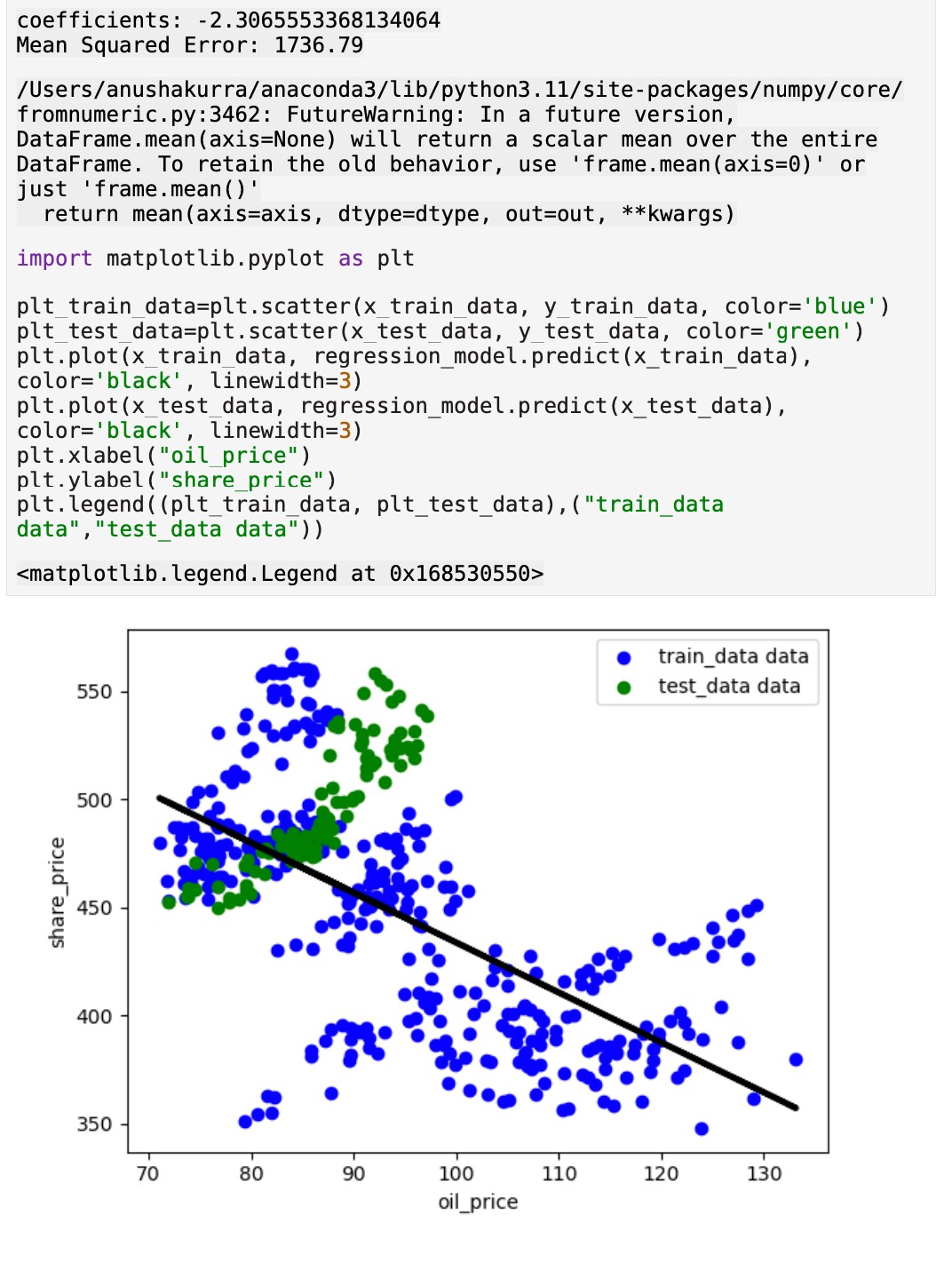




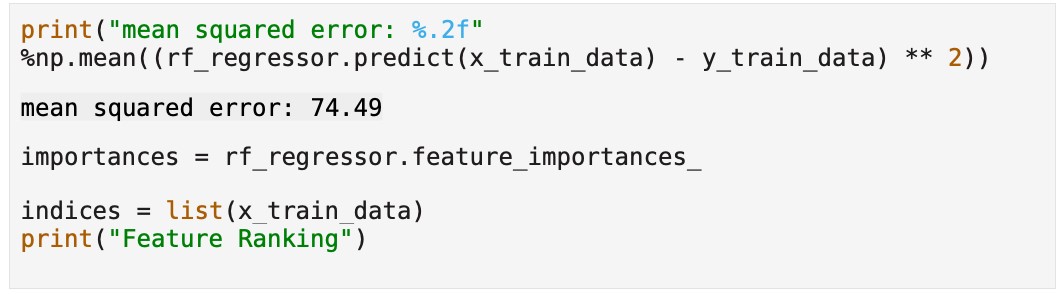
Simply by making predictions using the training data, the model seems to be rather accurate. very likely with a good deal of "overfitting" done to it. There are a lot of options to tweak, but one of the most important ones is the maximum depth. This will give the impression that there is more depth to trees.

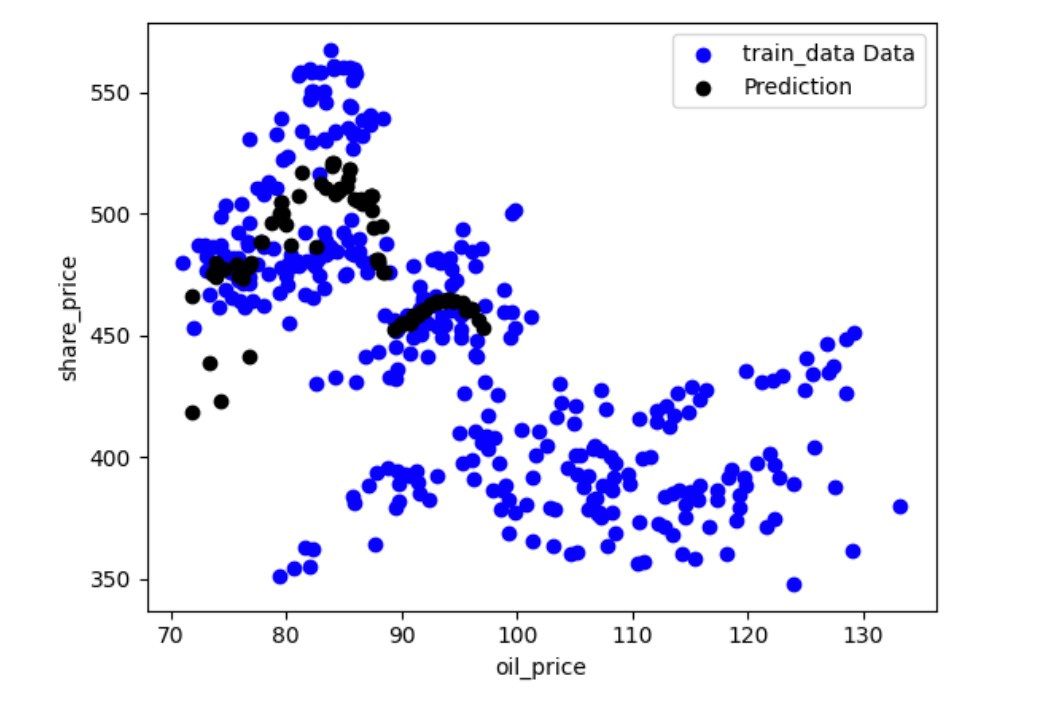


Linear regression results:



Random forest regressor results:





## Advantages and Limitations of the Approach

The random forest regressor results in a low MSE, leading to improved accuracy in forecasting oil stock prices. Traders and brokers have the potential to make a profit by buying oil stocks ahead of time. The potential for this to become the next phase in trading is a major threat to traditional traders. Financial analysts and forecasters find Python popular due to its user-friendly nature, wide range of libraries and tools, and capability to handle large datasets. Using Python for stock research and forecasting offers multiple benefits. Its extensive library support, encompassing Pandas, NumPy, and Matplotlib, simplifies the analysis of large stock market data sets. You can utilize Python's machine learning libraries like Scikit-learn and TensorFlow to develop models for predicting stock market trends. Seaborn and Polly are among the many Python modules that simplify the creation of compelling and informative stock market visualizations.

On the other hand, using Python for stock research and forecasting has its limitations. Newcomers without a background in computer science may find learning Python to be challenging. Predicting future stock prices using Python models may not always be reliable due to the inherent difficulty of the task. The accuracy of stock research and prediction can be affected by data quality constraints, such as missing or faulty data. To sum up, Python offers a powerful toolkit for analyzing and predicting stocks, but it's important to acknowledge its limitations. It's advisable to use other methodologies alongside Python for more accurate forecasting.

**Practical Applications of the Approach**

One example of Python's usefulness is in stock market research and forecasting. Python can be utilized to analyze and visualize historical stock price data, uncovering intriguing patterns and trends. Examples of this include: Nonetheless, without considering external factors and depending only on machine learning algorithms, it could result in significant financial losses for individuals. One can use libraries like Pandas, Matplotlib, numpy for this task. Python is capable of creating and testing trading strategies by utilizing historical stock price data. Some libraries that may be utilized for this include NumPy, Pandas, and Scikit-learn. Python has the capability to construct predictive models for forecasting future stock prices. There are several machine learning techniques such as linear regression, decision trees, and random forests that could be beneficial in addressing this issue.

**Conclusions:**

The use of supervised machine learning for oil and gas company share price prediction is demonstrated in this project report. The dataset utilized consists of the stock prices of four oil and gas firms combined with the prices of the Brent index oil. The two machine learning algorithms that are used are random forest regression and linear regression. The linear regression model was used to create a basic share price prediction model. The algorithm tried to forecast stock prices for the previous 100 days of data after being trained using current data. The model's mean square error of 2718.11 indicated that the results were not particularly accurate. Next, a more advanced model for share price forecasting was constructed using the random forest regression model. The algorithm tried to forecast stock prices for the previous 100 days of data after being trained using current data. With a mean square error of 911.12, the model's results were much more accurate than those of the linear regression model. The oil price was found to be the most significant feature in the model, with the share prices of the other oil and gas businesses coming in second. All things considered, this project showed how to forecast oil and gas company share values using supervised machine learning. The random forest model properly identified the price of oil as the most significant feature, outperforming the linear regression model in terms of accuracy.

In comparison to IEEE paper our values are different because we have taken different dataset. But we have accomplished what’s needed, which is finding a better model that is Random

Forest regressor.

References:

1. Peng, Zhihao. "Stocks analysis and prediction using big data analytics." *2019 international conference on intelligent transportation, big data & smart City (ICITBS)*. IEEE, 2019.

1. *USO - USCF Investments. (n.d.)*, [online] Available:

http://www.uscfinvestments.com/uso.

1. M. C. Angadi and A. P. Kulkarni, "Time Series Data Analysis for Stock Market Prediction Using Data Mining Techniques with R", *International Journal of Advanced Research in Computer Science*, vol. 6, no. 6, 2015.

4.G. V. Attigeri, P. MM, R. M. Pai and A. Nayak, "Stock market prediction: A big data approach", *TENCON 2015–2015 IEEE Region 10 Conference*, pp. 1-5, 2015.

5. Evangelos Triantaphyllou, C.-T.L., Development and evaluation of five fuzzy multiattribute decision-making methods. International Journal of Approximate Reasoning 1996. 14(4): p. 281-310.