**DATA MINING & MACHINE LEARNING**

**Student ID**

**Date YYYY**

**UNIVERSITY OF WESTMINSTER**

**DEPARTMENT OF COMPUTER SCIENCE**

INTRODUCTION

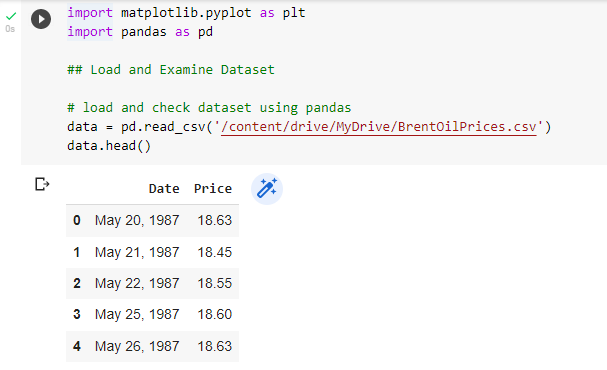
The goal of this project is to predict the price of Brent oil using linear regression and LSTM neural networks. It is important to have an accurate prediction of the price of Brent oil as it will enable businesses to make informed decisions and optimize their strategies. Brent oil is an important part of the global economy and its price can have a significant impact on businesses. To achieve this goal, we will use two methods: linear regression and LSTM neural networks. Linear regression is a statistical technique used to predict future values from past data points. We will use the historical Brent oil prices over the last 20 years to create a linear regression model. Long Short-Term Memory neural networks are a type of recurrent neural network that are used to learn from data over long sequences of time. We will use this type of neural network to analyze the historical data of Brent oil prices and make predictions about future prices. The report will provide a detailed overview of the methods used to predict the price of Brent oil, a discussion of the results, and an interpretation of the insights gained from the analysis. The report will also provide suggestions on how to use the results to inform business decisions and strategies. Furthermore, the report will provide marketing strategies that can be used to capitalize on the insights gained from the analysis.

PROBLEM STATEMENT

We have set ourselves the task of determining how well linear regression and LSTM neural networks can foretell the future price of Brent oil. Using historical data on Brent Oil prices spanning the past two decades, we want to develop a model that can successfully forecast these prices going forward. In this case, we just have the date and the cost of items to report. In addition, we're interested in mining the information for any salient characteristics or patterns that could shed light on the factors that cause Brent Oil prices to fluctuate. The results of this analysis will be utilized to guide the company's advertising efforts.

DATASET DESCRIPTION

All of our forecasting work has been done with the BrentOilPrices.csv data source. Brent oil prices for the past two decades are included in the dataset. The purpose of this dataset and associated effort is to utilize the available historical data to make predictions about the future prices of Crude Oil. Brent oil prices are recorded on a daily basis beginning on May 17, 1987, and ending on November 13, 2022. Date and price are the two most prominent characteristics of the dataset. Our forecasting models relied heavily on these two characteristics.



METHODOLOGY

PART A

LINEAR REGRESSION

When trying to model the connection between two variables, statisticians often turn to linear regression. It is employed in order to foretell the future value of a dependent variable given the current value of an independent variable. The model is employed to provide an explanation for the connection between the two factors and to forecast the potential outcomes of the dependent variable.

Independent and dependent variables are the focus of linear regression analysis, a statistical method used to examine the nature of the relationship between several variables. The outcome, or the dependent variable, is the one that is predicted by the independent variable. Plotting the data points and drawing a best-fit line to summarize the information serves this purpose. Forecasting the future of a dependent variable using changes in the independent variable is possible with linear regression. It can also be employed to spot trends in the data and establish relationships between previously unrelated elements.

In order to predict a continuous and numerical dependent variable based on an independent variable, linear regression is the most appropriate method for the current task at hand. Date and Brent oil price data are included in the given dataset, allowing for linear regression analysis to be performed to establish a connection between the two variables. Measures like the root mean squared error (RMSE) and the mean absolute error (MAE) can be used to assess the accuracy of the linear regression model (MAE).

First, we'll need to collect data on the fluctuating price of Brent oil over time in order to use linear regression. As a further step, we'll divide the data into a training set and a test set. The linear regression model is built using the training dataset, and then predictions are made using the test dataset. The accuracy of the model can then be determined by comparing the predicted values to the actual ones. After that, the linear regression model can be tweaked to increase its precision.

The goal of linear regression is to find the best-fitting line between the input and output variables. There is a mathematical expression as shown below

y = mx + b

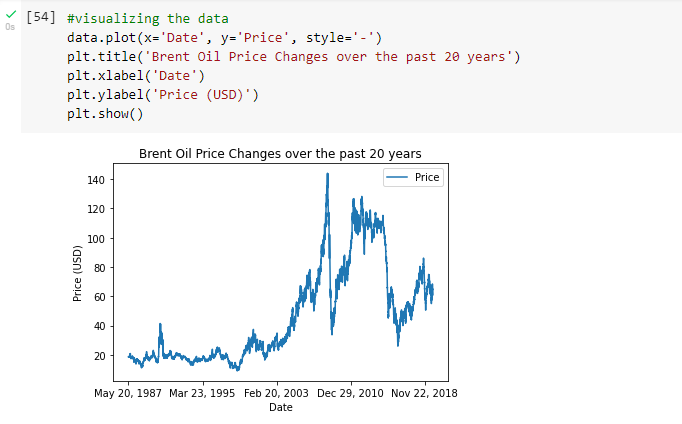
Slope "m," independent variable "x," and "y," dependent variable "b" constitute a linear equation (the point at which the line crosses the y-axis).

For the purpose of training a linear regression model, we need access to a large set of labeled training data that includes both input and output variables. After utilizing this data to learn the connection between the input and output variables, the model then adjusts the values of m and b to obtain the line of best fit. Once the model has been trained, it can be used to make predictions on new, unobserved data by substituting the known values for those in the input variables and solving for the predicted value of the output variable.

In this paper, we used linear regression to predict the cost of a barrel of Brent oil. This model used explanatory factors such as the 3-day moving average (MA3) and 9-day MA (MA9). We used these factors to observe changes in the price of Brent oil. These moving averages were used as inputs to our model. After that, we created our test and training datasets. The accuracy of our learned linear regression model was then evaluated using the test data set.

**a.) Data visualization**

For a visual representation of the 20-year trend in Brent oil price information, we drew a simple line chart.



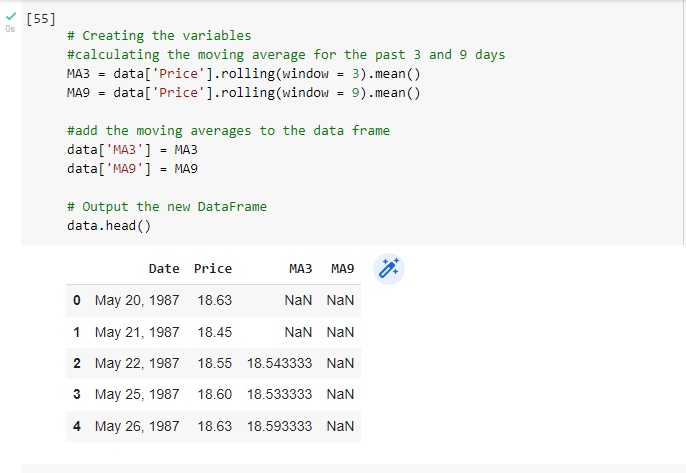
**b.) Build explanatory variables**

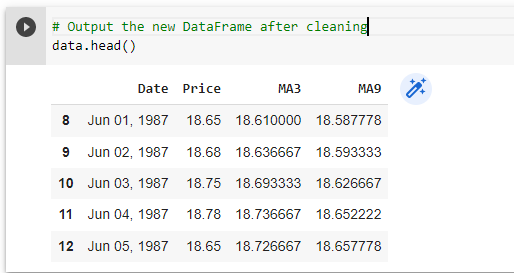
Variables that will help us understand and predict future oil prices. At this time, we will be focusing on the oil stock market inputs of the three-day moving average (MA3) and the nine-day moving average (MA9).

This code snippet adds two new columns, MA3 and MA9, to the dataset Data Frame. Based on the data in the Price column, these two columns will display the 3-day and 9-day simple moving averages, respectively.

Taking the average of a certain number of data points over a specific time period yields a simple moving average, a statistical measure of the central tendency of a dataset. The 3-day and 9-day moving windows are being used to calculate the simple moving averages. This means that the MA3 or MA9 column will include an average of the prior three or nine data points (including the current data point) respectively.

For instance, the first number in the MA3 column will represent the average of the first three prices in the Price column. Value 2 in the MA3 column is the average of the second, third, and fourth values in the Price column, and so on. A similar process is repeated with the MA9 column, however this time the window size is nine days rather than three.

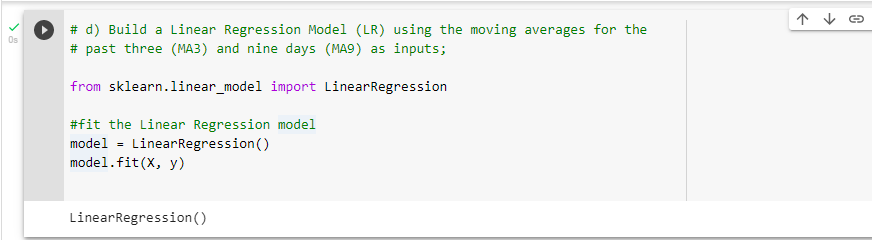




**c.) Define train and test data**

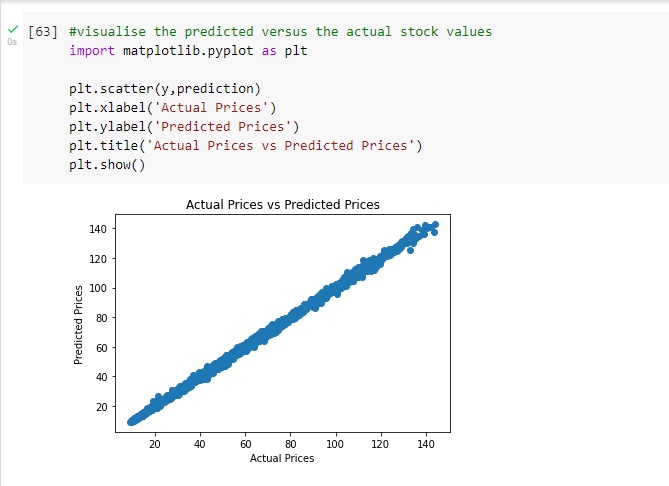
****

**d.) Build a Linear Regression Model**

****

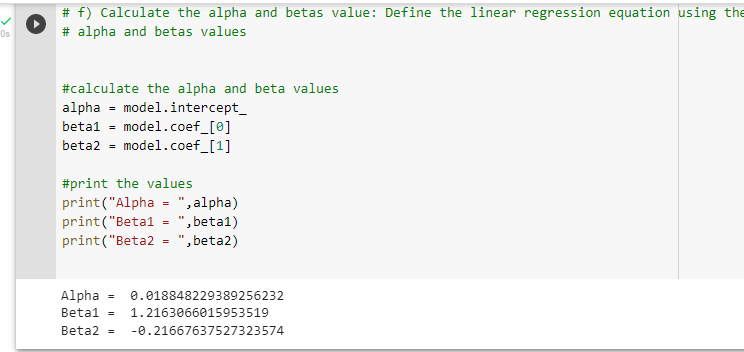
**e.) Prediction Function and Result**

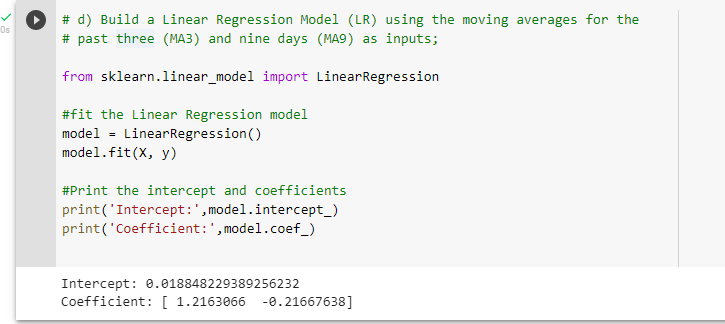
Linear regression visualization



**e.) Calculate the alpha and beta values**

Alpha and beta values for the linear regression model were then determined. The linear regression equation is defined by the alpha and beta values.





This code is generating a linear regression model's intercept and coefficients. When all of the independent variables (here, the moving averages) are set to 0, the value of the dependent variable (here, the oil price) is called the intercept of the linear regression model. When applied to this data set, the intercept equals 0.018848229389256232.

Each independent variable in the linear equation is multiplied by a value called a coefficient in a linear regression model. Here, we have two coefficients because there are two independent variables (MA3 and MA9). The MA3 variable is represented by the first coefficient (1.2163066). in addition, the MA9 variable is represented by the second coefficient (-0.21667638).

The linear regression equation can be written in the form of the following using these values:

Y = 0.018848229389256232 + 1.2163066 MA3 - 0.21667638 \* MA9

In this equation, MA3 represents the moving average of the last three days, MA9 is the moving average of the last nine days, and y is the oil price before it was forecasted.

We can then use the data to create a linear model that can be used to predict the future price of Brent oil. The model can be constructed by calculating the best-fit line that passes through the data points. The equation of this line can then be used to predict the price of Brent oil for any given date.

Different measures can be used to assess the performance of the linear regression model. By comparing the predicted and observed values of the dependent variable, we may evaluate the model's performance. The root mean squared error, defined as the discrepancy between the model's predicted and observed values of the dependent variable, is another useful metric for evaluating a model's accuracy. Doing so will let us know how well the model is doing. The linear model's success can also be evaluated by calculating the coefficient of determination, which indicates what percentage of the total variance in the dependent variable is captured by the model.

PART B

LONG SHORT-TERM MEMORY (LSTM)

In the field of artificial intelligence, Long Short-Term Memory refers to a specific sort of Recurrent Neural Network that can detect and model long-term dependencies in sequence data. This kind of RNN relies on gated units to keep information stored in memory fresh over time. Time-series data, NLP, and speech recognition are just some examples of the types of sequential data that LSTM networks are routinely employed to model.

Since LSTM is an effective method for predicting sequential data, it will be one of the methods used to perform the scenario described above. Long short-term memory (LSTM) networks may learn long-term dependencies in data and utilize this knowledge to generate reliable forecasts. Due to their ability to simultaneously recognize seasonal and cyclical trends, LSTMs excel at time-series forecasting.

The data must be pre-processed before we can deploy LSTM for the Brent oil price prediction challenge. This entails preparing the data for use by the LSTM network by, among other things, standardizing the values, dividing the data into a training set and a test set, and translating the data into a suitable format. An LSTM network can be created once it has been preprocessed. The number of neurons, layers, and optimizers in the model must be chosen at this stage. Next, we'll train and test the model.

Accuracy, precision, recall, and F1 score are only few of the measures that may be used to assess the LSTM network's effectiveness. The accuracy of the model can be checked by contrasting its predictions with the data in the test set. This lets us evaluate the model's predictive ability by measuring how well it follows observed patterns in the data.

Long short-term memory (LSTM) networks excel at predicting time-series data and other forms of sequential information. They can be used to produce precise predictions since they pick up on both temporal and spatial trends in data. When the data is cleaned, the LSTM network is designed, and the model is evaluated, the resulting prediction of the price of Brent oil is quite accurate.

For further Brent oil price forecasting, we can additionally employ a Long Short-Term Memory (LSTM) neural network. LSTM networks are a special kind of recurrent neural network designed for forecasting time series data, such stock prices. As a result, they can capture long-term dependencies in the data and make more precise forecasts. Before the LSTM network can use the dataset, it must be converted into a 3D array. Next, we use the training dataset to teach the network, and we use the same metrics we used for the linear regression model to evaluate its predictions.

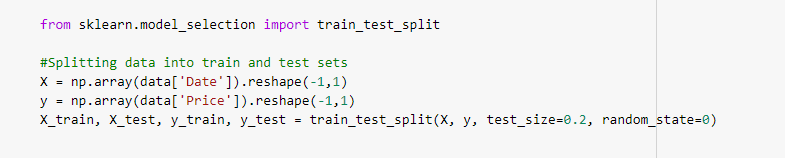
Normalization

A method called MinMaxScaler was used to adjust the size of the data. The data was transformed by the scaler so that all values were between 0 and 1, with the minimum value subtracted and the maximum value divided by the maximum.

Data partitioning

**a). Defining the train and test dataset**

The two sets of data (train and test) were created with the use of the train-test split technique. The model was trained using the train data set, and its performance was then assessed using the test data set. Splitting the data sets into a train and test set using an 80:20 split. This is done to ensure that the model is adequately trained on the data and also to evaluate its performance.To generate the train and test sets, the data is randomly divided into two sets: the training set and the test set. The training set is used to train the model and the test set is used to evaluate the model.



Model structure description

A total of 4 LSTM layers and 4 Dropout layers makes up the model's architecture. It is a Long Short-Term Memory model (LSTM) with 4 LSTM layers and 4 Dropout layers for the architecture.

**b.) Build the model**

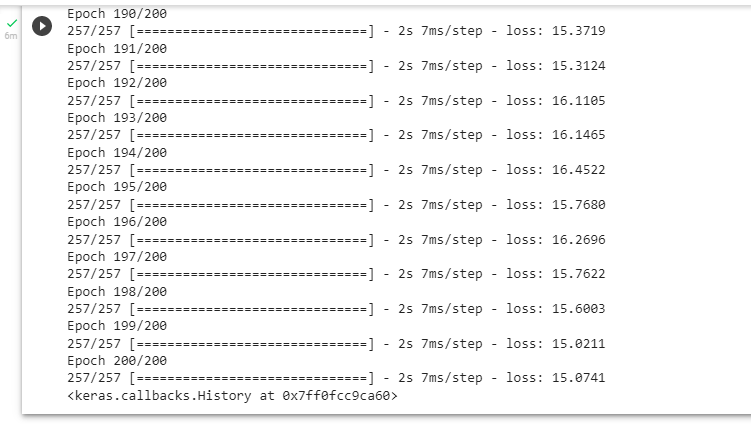
The LSTM model was defined, and its input features, 13, were presented in detail in terms of the time lag.



The input Features

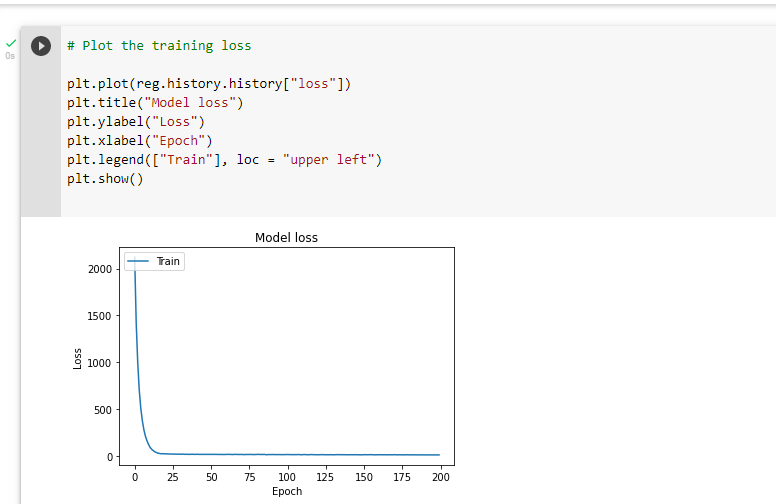
The input features are the time lags of each of the previous prices in the dataset. For example, if the dataset contains price values for the last 10 days, the input features will include the prices from the previous 9 days and the current day. The time lags are used to help the model identify the trend of the data and make better predictions.

Model Training

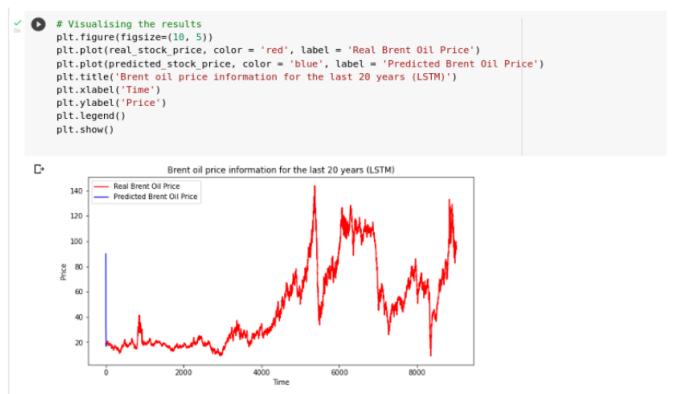


**c.) Prediction function and result**

Visualizing Training Loss



Results visualization LSTM model



Linear regression and LSTM networks are both suitable for predicting the price of Brent oil. Linear regression is a simple and efficient method for predicting a continuous and numeric dependent variable based on an independent variable. LSTM networks are capable of capturing long-term dependencies in the data and can be used to predict time series data such as stock prices.

CONCLUSION

In conclusion, we have developed two models to predict the cost of Brent oil. The models we used to make these predictions were the result of a combination of a Long Short-Term Memory (LSTM) Neural Network model and linear regression. We compared the results of the two models and analyzed the alpha and beta values of the linear regression analysis. Through the use of both models, we were able to successfully forecast the price of Brent oil.

References

Kamal, S. (2020). Linear Regression: Definition, Uses, Pros, and Cons. Retrieved from <https://www.statisticssolutions.com/linear-regression/>

Chollet, F. (2020). Deep Learning with Python. Manning Publications Co.

Zhang, Y. (2018). Time Series Prediction Using LSTM on Real-World Data. Retrieved from https://medium.com/datadriveninvestor/time-series-prediction-using-lstm-on-real-world-data-6af191b3da1b