Dataset

The Ford dataset, sourced from the Yahoo Finance library, comprises a comprehensive set of attributes essential for stock market analysis. The dataset contains the following features:

1. **Open:** The opening price of Ford stock for a particular trading day.
2. **High:** The highest price reached by Ford stock during the trading day.
3. **Low:** The lowest price reached by Ford stock during the trading day.
4. **Close:** The closing price of Ford stock for the trading day.
5. **Volume:** The total number of Ford stock shares traded during the trading day.
6. **Dividends:** Any dividends declared by Ford during the trading period.
7. **Stock Splits:** Any instances of stock splits that occurred during the trading period.
8. **Date:** The date corresponding to the stock market trading day.

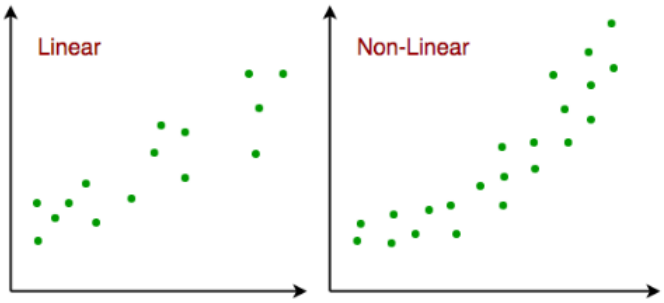
These attributes provide a comprehensive view of Ford's stock performance over time. The relationship between these attributes is essential for understanding various aspects of Ford's stock behavior, including price fluctuations, trading volume, dividend payouts, and stock split occurrences. Analyzing these attributes collectively enables investors and analysts to gain insights into Ford's market trends, volatility, and overall financial health.

Exploring the relationships between these attributes can uncover patterns, correlations, and trends that influence Ford's stock performance, facilitating informed investment decisions and strategic planning for stakeholders in the financial markets.

Literature Review

“Machine learning (ML) is a branch of Artificial Intelligence(AI) and computer science that focuses on the use of data and algorithms to imitate the way that humans learn and behave to improve its accuracy gradually. (IBM)” Machine learning is automated, and the computer program is “learned” from data, meaning it automatically makes improvements based on continual incoming data. With machine learning, you can design methods applicable to various practical applications in various areas because they can represent an arbitrary data set. Machine learning’s major objective is to classify data based on models and develop predictions based on these models(Lovell, 2023). We will use Python machine-learning statistical techniques such as Linear Regression, Logistic Regression, Decision Trees, and Random Forests to help us predict stock on Ford.

Linear Regression is a statistical method that allows a programmer to model relationships between “target” and “predictor” variables. Linear regression predicts a dependent variable or a target based on the independent variables. It uses the relationships between the data points to draw a straight line through them, and the line can be used to predict future values.

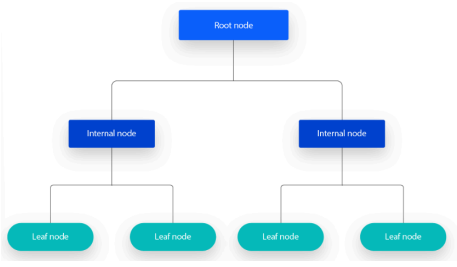


The linear regression model makes the following assumptions when applied to datasets. It first assumes a linear relationship between the target and the predictor variable. For example, the figure to the right shows the difference between a linear and non-linear graph. You can test your data by graphing a scatter plot to observe the

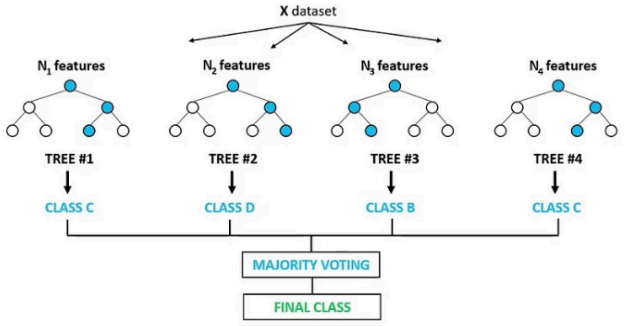
relationships. The linear regression method also assumes a modicum of multicollinearity in the data. If this occurs, then it proves that the independent variables are not independent of each other. It also assumes there are no outliers in the data; outliers can negatively affect the data's results. All in all, regression is a technique used to estimate relationships between datasets.

Logistic Regression is a classification algorithm that describes data and explains the relationship between one dependent(binary) variable and one or more independent variables. It uses the relationship to predict the value of one based on the other. The algorithm's output is either 0(failure) or 1(success). "Logistic regression models can process large volumes of data at high speed because they require less computational capacity, such as memory and processing power. This makes them ideal for organizations starting with ML projects to gain quick wins(AWS)." The finance industry famously uses logistic regression to analyze transaction fraud and to assess applications for loans, insurance, etc. Logistic regression works for these problems because they return discrete outcomes such as "high risk" or "low risk".

"Decision Tree classification algorithm is a tree-based model that consists of internal nodes, branches, and leaves. The internal nodes represent the attributes of the data, the branches represent the decisions or rules, and the leaves represent the outcomes or predictions(Medium)." Decision trees begin with a root node, which branches out to internal nodes called decision nodes. Decision trees make decisions based on values from the features called homogeneous subsets, and these subsets are denoted by leaf nodes. The leaf nodes represent all possible outcomes of the dataset. The figure to the right shows an example of what a decision tree looks like. At each node, the decision tree splits the data based on the attributes that give the highest data gain until all the data points of the same leaf are categorized into the same class.

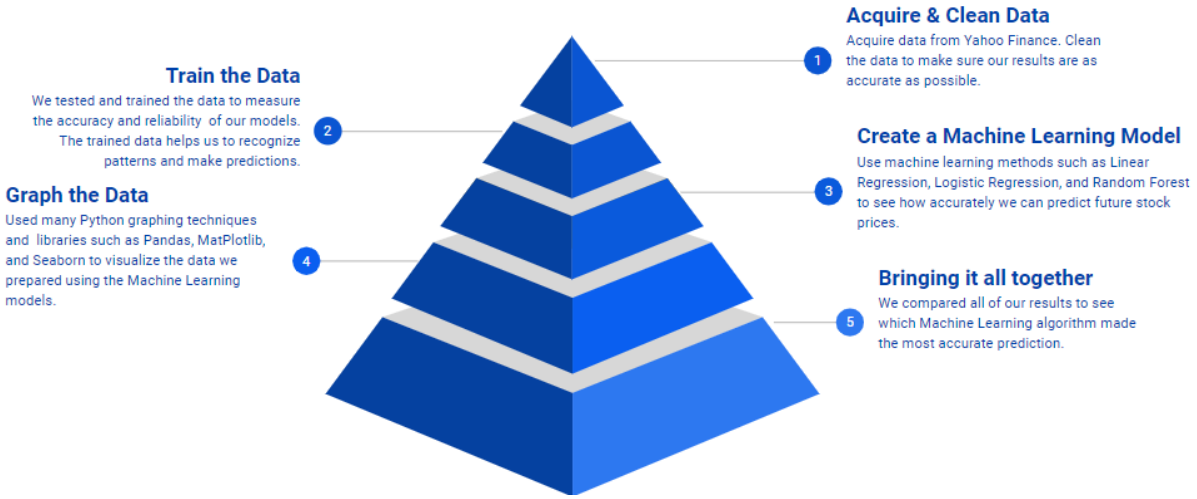


“Random Forest is an ensemble learning method that combines the predictions from multiple decision trees to produce a more accurate and stable prediction. It is a type of supervised learning algorithm that can be used for both classification and regression tasks(GeeksforGeeks).” This method uses multiple decision trees rather than one decision tree to determine the output by using a “majority vote”. Random Forest contains trees(a tree represents a decision tree). We get these trees by taking a dataset and dividing the dataset into batches or smaller trees that contain random data from the original dataset and build decision trees for each of them. The figure above shows an example of the Random Forest algorithm. From these multiple trees, you create a “forest” of trees. Fun fact: This is where the idea of its name was developed. Once the user trains the data for what they want to predict, the trees would each give their result. Based on the tree’s result, the majority vote from all trees determines the final output.



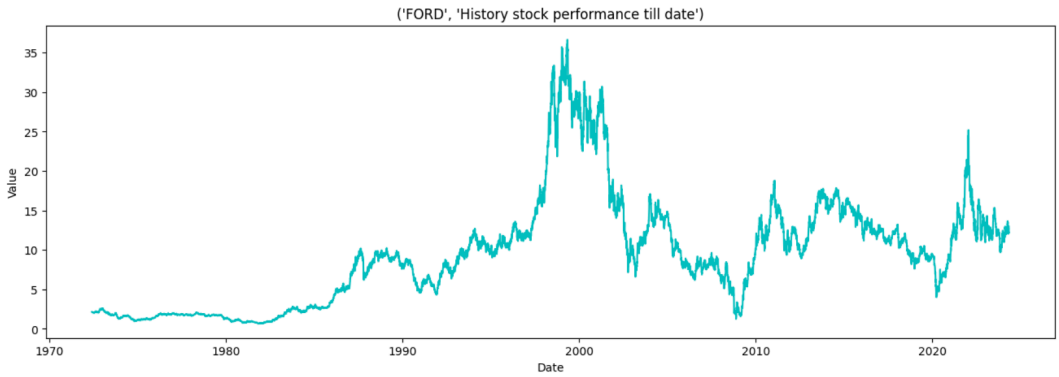
There are many such algorithms present that can help a user with stock prediction. As you can see, Linear Regression, Logistic Regression, Decision Trees, and Random Forests are some of the few but widely used and known algorithms used to help one with exploratory data analysis and stock prediction. Later in our results, you will observe how we used these algorithms to assist us with Ford stock prediction.

Architecture/Methodology

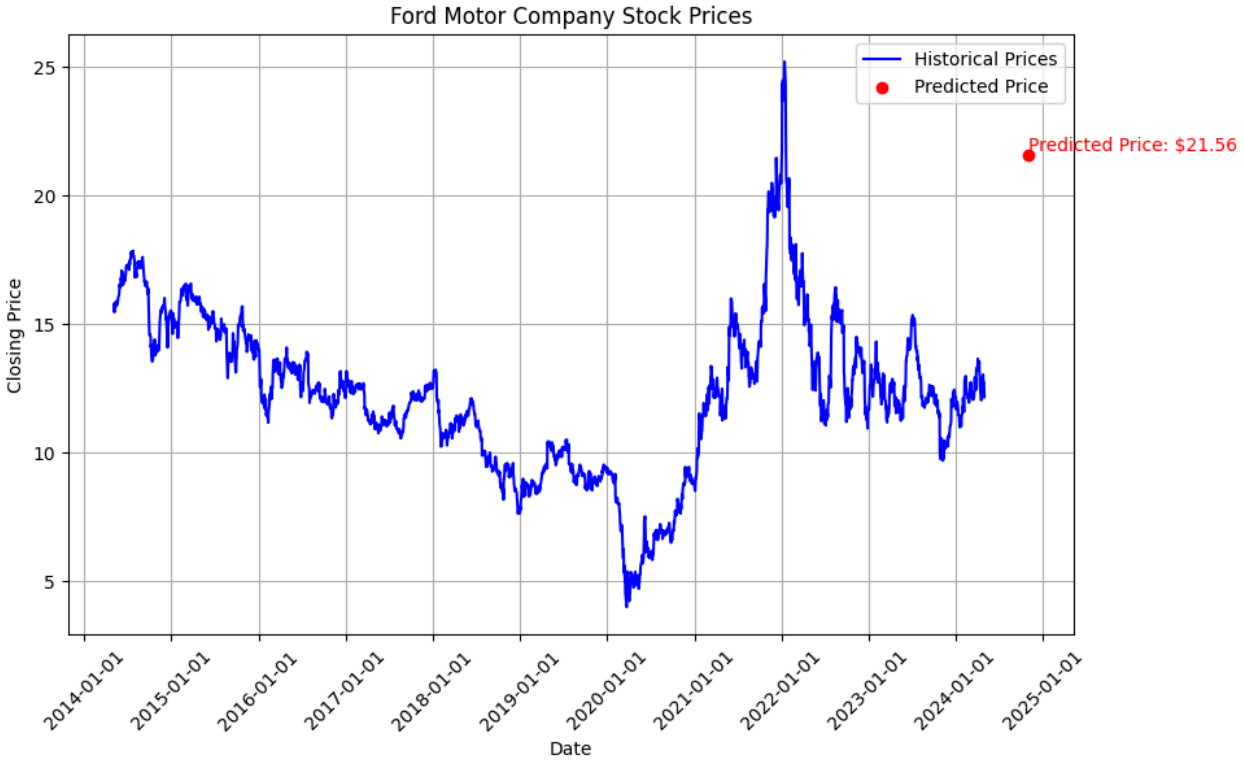


Results- Ford Motor Co. (FORD)

Ford Motor Co. is an American multinational company that manufactures, distributes, and sells automobiles. We used the YahooFinance library to import Ford's historical stock data. We then cleaned and observed the stock data to remove any inconsistencies that may negatively alter the data. We graphed Ford's historical stock data from the beginning to current to observe the trend of the company's stock. As you can see, based on the graph below, Ford's stock price is pretty consistent currently. It doesn't show any volatility within its data. Ford also had its highest stock price between 2004 and 2008, when its stock price was over $25 USD.

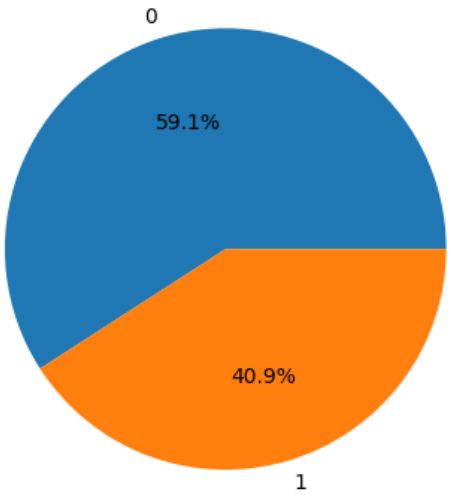


As mentioned above, we used Yahoo Finance to import the historical stock data for Ford Motor Co. After cleaning the data, we split the data into training and test datasets. The test dataset was 15% of the overall dataset. We used Linear regression () because we wanted to find a relationship between our independent and dependent variables to be able to predict stock data prices. We then measured the Mean Absolute Error(MAE) to measure the average amount of mistakes within our prediction. Our MAE was 0.07%, meaning there were almost no errors in our predicted and actual values. We also measured the root mean square error (RMSE), which gives you the average difference between our predicted and actual values, and our result was 0.10%. Both results were less than 1%, meaning we were almost completely accurate with our predicted and actual results. You can see from the graphs below our Actual stock price results compared to our Predicted stock price results. We believe Linear Regression gives a pretty good estimation of predicted stock prices using our actual stock price dataset.



This section utilized logistic regression to analyze Ford stock data sourced from the Yahoo Finance library. The code provided demonstrates the data preparation process, training the model, and evaluating its performance.

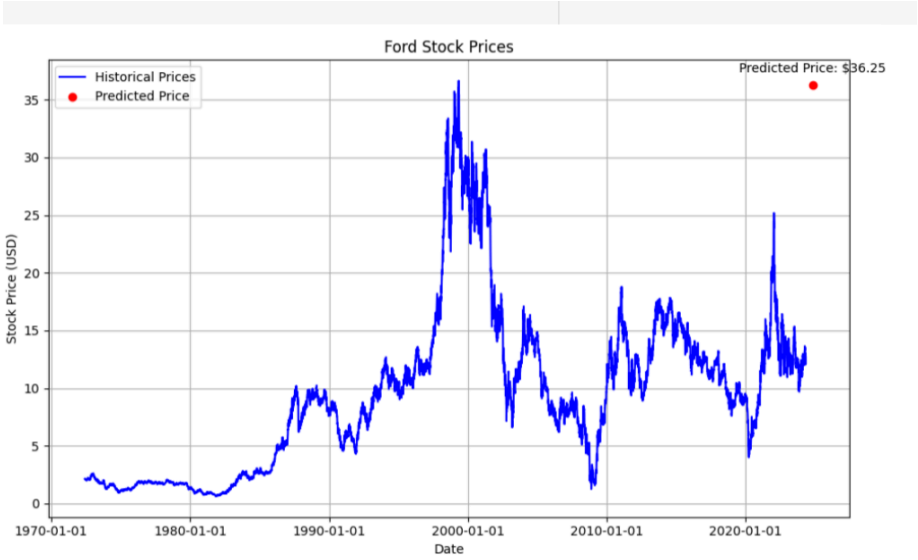
First, we engineered additional columns, 'open-close' and 'low-high', to capture relevant information regarding price differentials within each trading day. Additionally, a 'target' column was created to classify whether buying the stock is optimal based on the closing price trend. To ensure the target variable is balanced, we visualized its distribution using a pie chart, showcasing the proportion of buy (1) and not buy (0) signals. Subsequently, we split the dataset into training and validation sets, selecting the engineered features ('open-close' and 'low-high') as predictors for the model. The data was normalized using standard scaling to ensure uniformity across features. Logistic regression was then applied as the predictive model. We evaluated its performance on the training and validation sets using the Receiver Operating Characteristic Area Under Curve (ROC AUC) score. The ROC AUC score measures the model's ability to discriminate between positive and negative signals, with a value closer to 1 indicating better performance.



The results indicate a training accuracy of approximately 57.3% and a validation accuracy of around 57.6%. While the model demonstrates some predictive capability, further refinement and feature engineering may be necessary to enhance its accuracy and reliability for real-world stock trading decisions.

We then ran the code to determine if the random forest algorithm is the correct technique to train the machine learning model. Upon reviewing the MSE (Mean Square Error) of 0.01246%, we can say that this is a reliable model and that the price prediction should accurately predict the real closing price in October 2024.

In the context of machine learning, feature scaling is often performed to standardize or normalize the features in the dataset. This is important because it ensures that features are on a similar scale, which can help improve the performance and convergence of many machine learning algorithms. Hence, we decided that for the random forest model we needed to normalize the dataset to get a more accurate model.



Per our random forest model, Ford's stock closing price on October 30, 2024, should be close to $36.25, which is around the stock's historic peak price of $2000.

Conclusions

Although we used the same dataset, and our statistical measures tell us that each of the trained models is reliable, we can clearly see that the model trained using linear regression tells us that the stock price is going to continue losing value and reach a predicted value of $21.56, while the model trained using the random forest technique tells us that Ford's stock price will go back to its historic peak price of around $36.

We can say that based on the statistical measures, both the linear regression model and the random forest model seem to be very reliable, but they both arrive at very different predicted prices for the selected stock. Therefore, we conclude that right now, we would need to go further into financial analysis, the company's debt structure, sales forecast, and vehicle consumer market to build a reliable financial model for Ford's stock price and determine which machine learning model gets closer to the forecasted price using the widely use financial forecasting technique.

Using the algorithms trained using three different techniques, we can conclude that the models are not 100% reliable and are a bit far from being a guide for investors. Nevertheless, with fine-tuning, we think the models can return accurate predictions, assuming there are no external geopolitical, natural disasters, etc., that may inadvertently affect the company's regular operations. In addition, our Logistic Regression indicated a training accuracy of approximately 57.3% and a validation accuracy of around 57.6%. While the model demonstrates some predictive capability, further refinement and feature engineering may be necessary to enhance its accuracy and reliability for real-world stock trading decisions.

In the last few months, Ford Motor Company has outpaced its competitors. We will have to wait and see where does it stock price closes in late October 2024; between the linear regression model and the random forest model, there is almost a $15 discrepancy, which allows us to confidently say that although the stock should gain

some value in the following months, we still need to work on this models so that two models trained with the same exact dataset don’t return very different results.