ML PROJECT PROPOSAL

PROJECT: NETFLIX STOCK PREDICTION USING RANDOM FOREST

PROBLEM STATEMENT

Netflix stock price prediction dataset is to predict the future stock price of Netflix based on historical data. The dataset contains daily stock prices of Netflix from 2002 to 2021, along with relevant financial indicators such as volume, market capitalization, and dividend yield.

The purpose here is to make a prediction of a continuous numerical value (the stock price) using a different set of input characteristics, making this a regression issue. For this purpose, the well-known machine learning method known as "Random Forest" might be quite useful.

The machine learning method Random Forest may be used for both regression and classification purposes. It is an ensemble learning technique that generates many decision trees during training and uses their average prediction to get a final prediction. By using Netflix's stock price history for training and making predictions for the future, Random Forest may be utilised to tackle this issue.

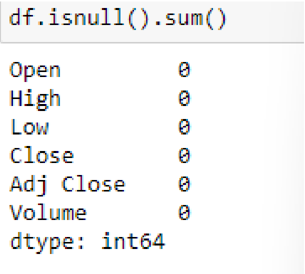
The result of the Random Forest method is the combination of decision trees built on independently-generated subsets of data. Since it can account for non-linear correlations between the characteristics and the objective variable, Random Forest is an effective technique for forecasting stock prices.

METHODS

Random forest is a popular machine learning algorithm that can be used for regression tasks like Netflix stock price prediction.

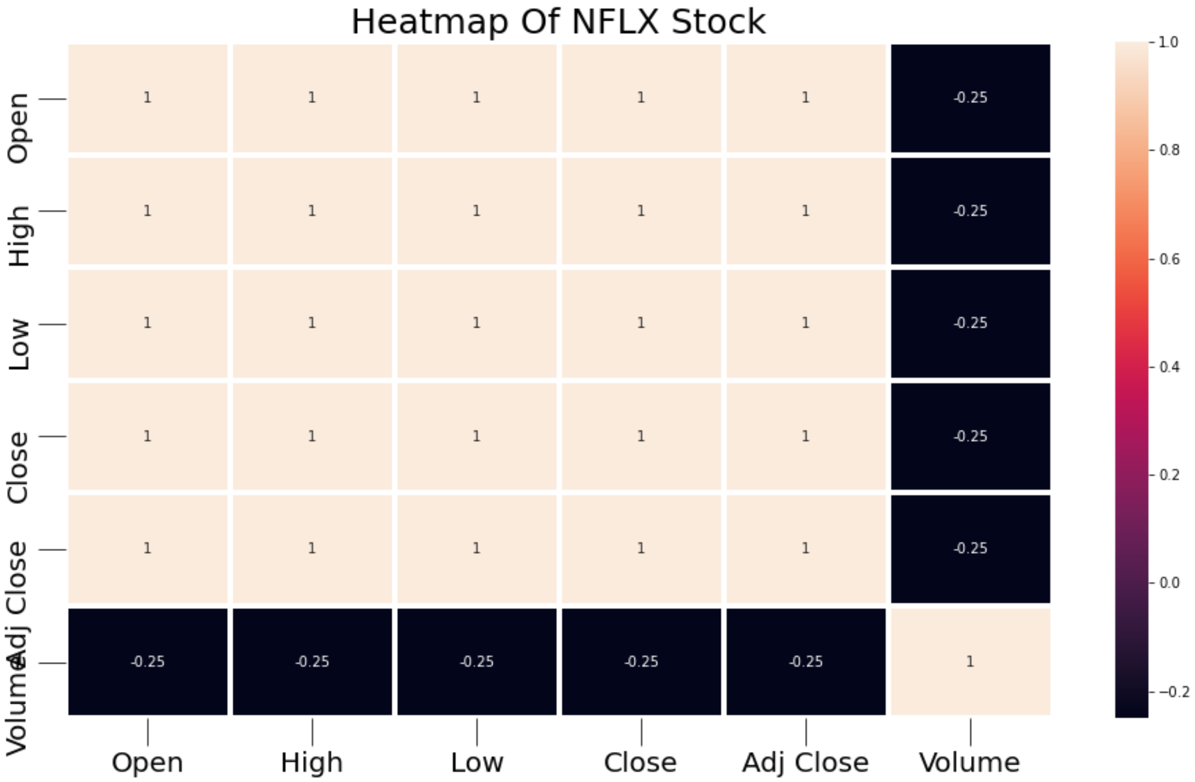
1. **Data Pre-processing**

   The first step is data pre-processing, which may entail, among other things, eliminating duplicates, reducing outliers, and scaling the data. In this dataset, there are no duplicates outliers and missing values as per the following output



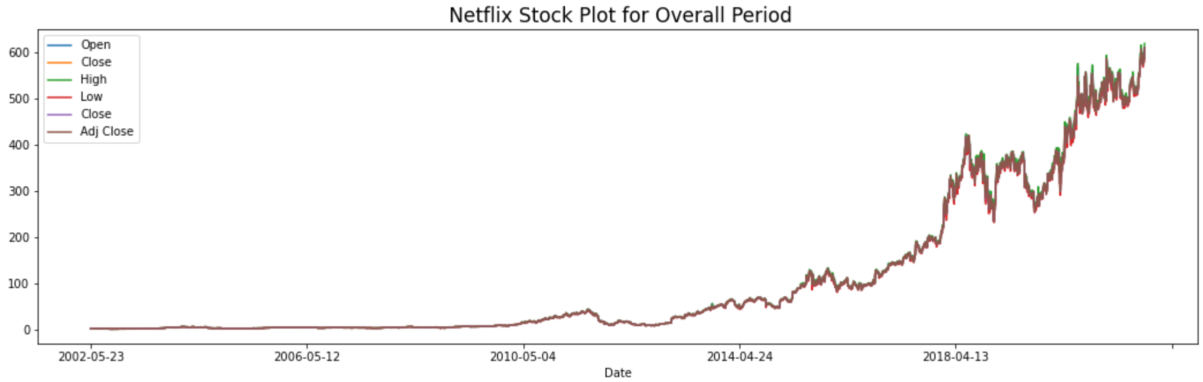
```
df.isnull().sum()

Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

   As per the following correlation plot, the interpretations are as follows:

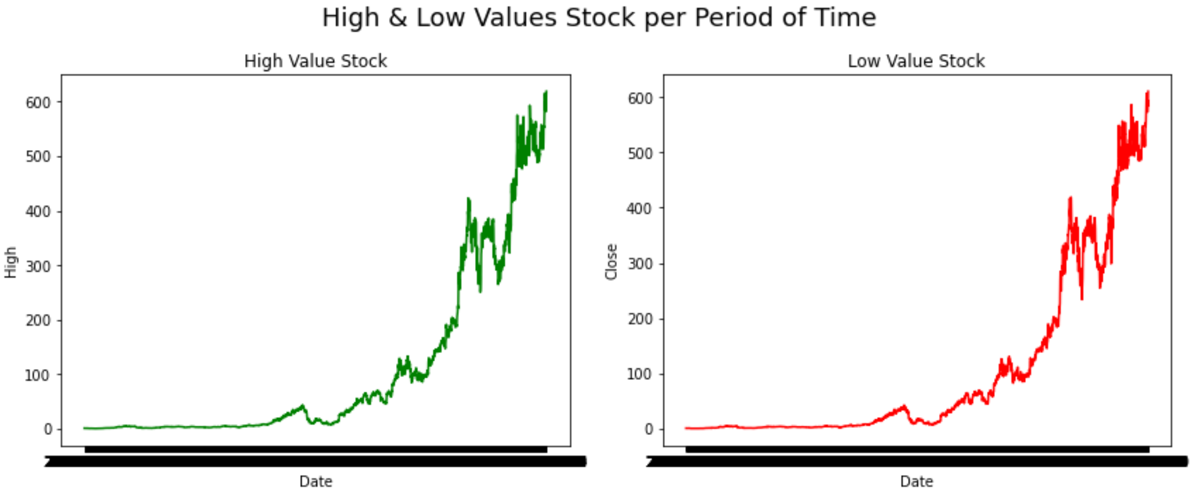


- Four daily indicators are used to monitor the price of Netflix shares: open, high, low, and close. Since an increase in one variable typically leads to an increase in the others, these variables are highly positively correlated. The crimson diagonal line in the correlation diagram highlights this pattern of correlation. However, these factors have little to no correlation with Volume.
- The Volume variable keeps note of the number of shares that trade hands on any given trading day. It has an inverse relationship with the market pricing. This finding suggests there is no significant correlation between trading volume and stock prices.
- The relationship between the high and low values is the strongest. This indicates that there is a positive relationship between the stock's closing price and its daily high, indicating that they move in tandem.
    - The positive association between Low and Open is weaker than the positive association between High and Close, but it nevertheless exists. This indicates that the initial price of the stock and its daily minimum have a similar trajectory.

2. Data Exploration

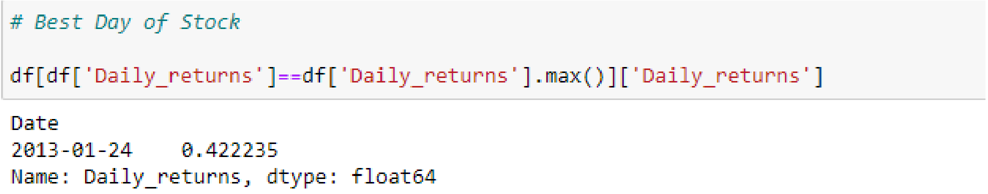


As per the above plot the Netflix stock has increased over the years and still has an increasing trend



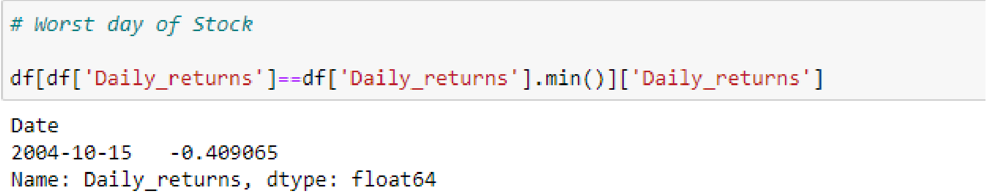
As per the above plot both low and high value stock has shown an increasing trend over the years and it still continues to grow

The worst and best day of stocks between the period of 2002 to 2021 can be seen from the following snippet of code and output



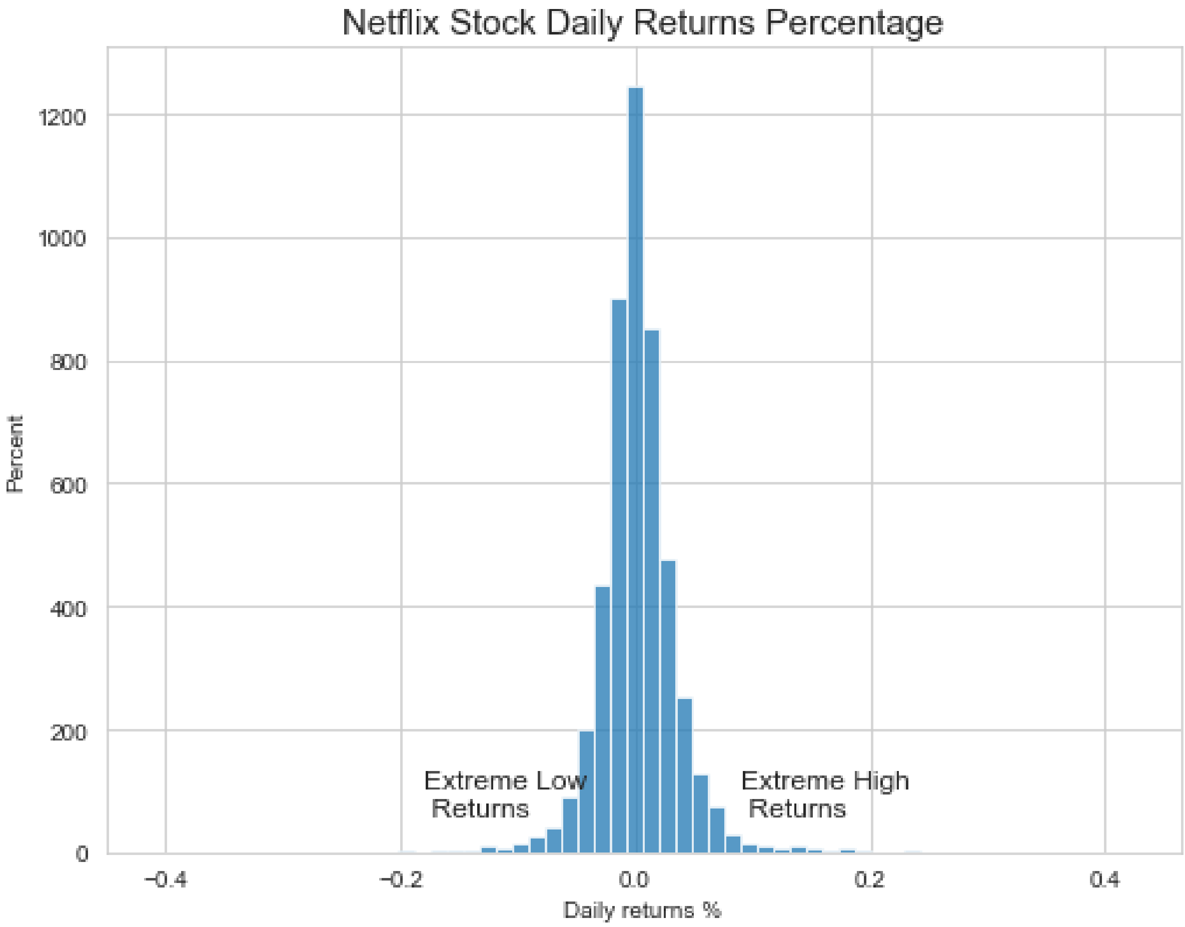
```
# Best Day of Stock

df[df['Daily_returns']==df['Daily_returns'].max()]['Daily_returns']
```

```
Date
2013-01-24    0.422235
Name: Daily_returns, dtype: float64
```



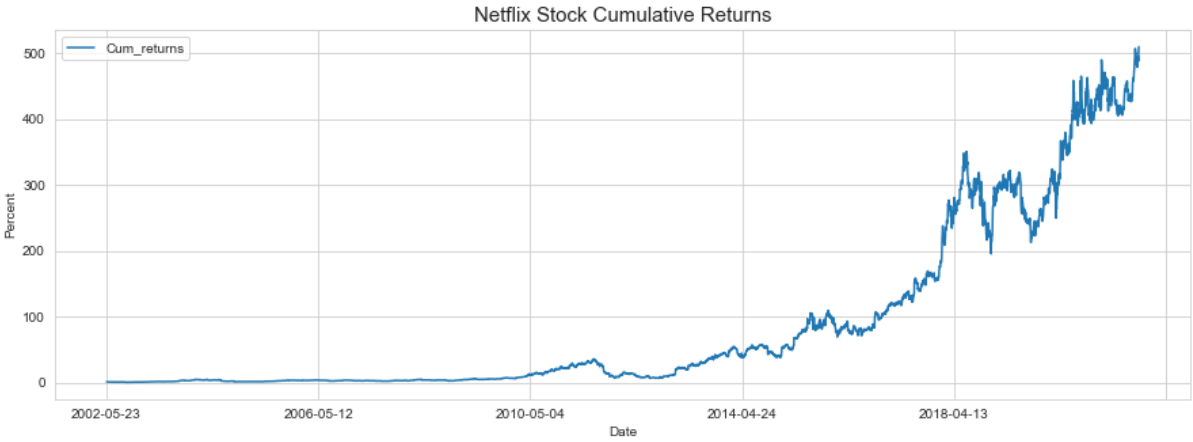
```
# Worst day of Stock

df[df['Daily_returns']==df['Daily_returns'].min()]['Daily_returns']
```

```
Date
2004-10-15   -0.409065
Name: Daily_returns, dtype: float64
```

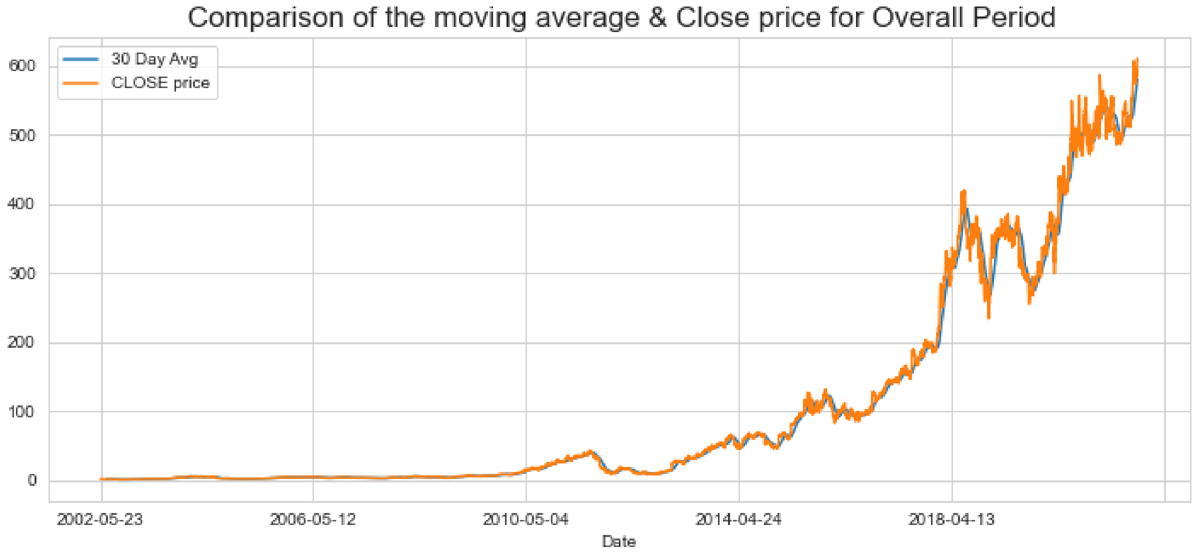


As per this plot, it can be said that the daily stock price of Netflix has a normal distribution. This information can be helpful in lot of ways, some of them are:

- If Netflix's daily stock prices follow a normal distribution, we can use statistical models that assume normality to generate price forecasts for the future.
- For assessing risk, data from a series of stock prices with a normal distribution may be useful. The standard deviation of a stock's volatility is a crucial metric for risk management.
- If daily fluctuations in Netflix's stock price incline towards a normal distribution, it may be indicative of an efficient market. If stock prices follow a normal distribution, as the efficient market hypothesis predicts, then the market is likely assimilating and pricing all relevant information effectively.



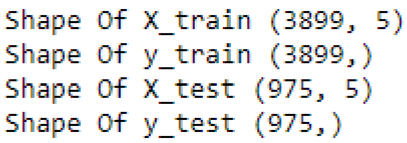
As per the above plot, the Netflix cumulative returns are also in increasing trend.



From the above plot, it can be said that the moving average of the stock price aligns with the close price for the period of 2002 to 2021.

3. Train-Test Split

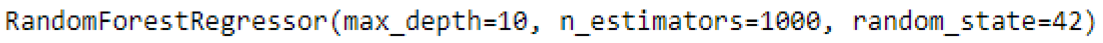
The dataset is split into train and test in 8:2 ratio i.e. 80% of the data has been used for training and remaining for the testing and validation of the model. The distribution of the data can be seen from the following snippet of output.



```
Shape Of X_train (3899, 5)
Shape Of y_train (3899,)
Shape Of X_test (975, 5)
Shape Of y_test (975,)
```

4. **Model**

   Next step was fitting the random forest model to the training data. And then this model will be tested against the test data to evaluate its performance. The hyperparameters was set as following:



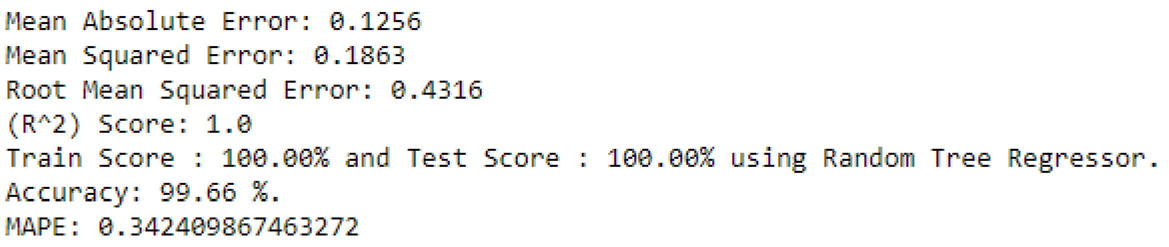
```
RandomForestRegressor(max_depth=10, n_estimators=1000, random_state=42)
```

EVALUATION CRITERIA

The evaluation criteria for predicting stock prices using Random Forest can includes the following metrics:

- The RMSE is calculated as the square root of the mean of the squared discrepancies between actual and anticipated stock values. It is a statistical measure of the average deviation of forecasts from reality. Lower levels of root-mean-squared error (RMSE) indicate enhanced performance.
- R2 indicates the proportion of the dependent variable's variability that can be explained by the model. The optimal score would be 1, whereas a score of 0 would indicate no progress. If R2 is high, performance is outstanding.
- MAE is the average of the absolute differences between actual and predicted stock prices. This statistic measures the average deviation of expectations from reality. The closer the MAE value is to zero, the greater the performance.
- Mean Absolute Percentage Error is the average of the absolute percentage mistakes between actual and anticipated stock prices (MAPE). It measures the deviation between the average prediction and the actual outcomes. In general, a lower MAPE value reflects more performance.

The following are the values for these metrics discussed above.

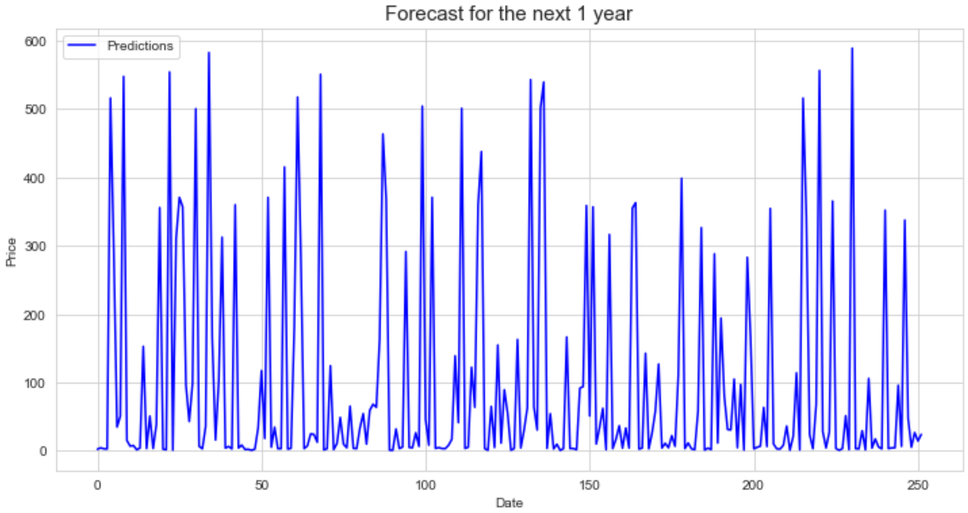


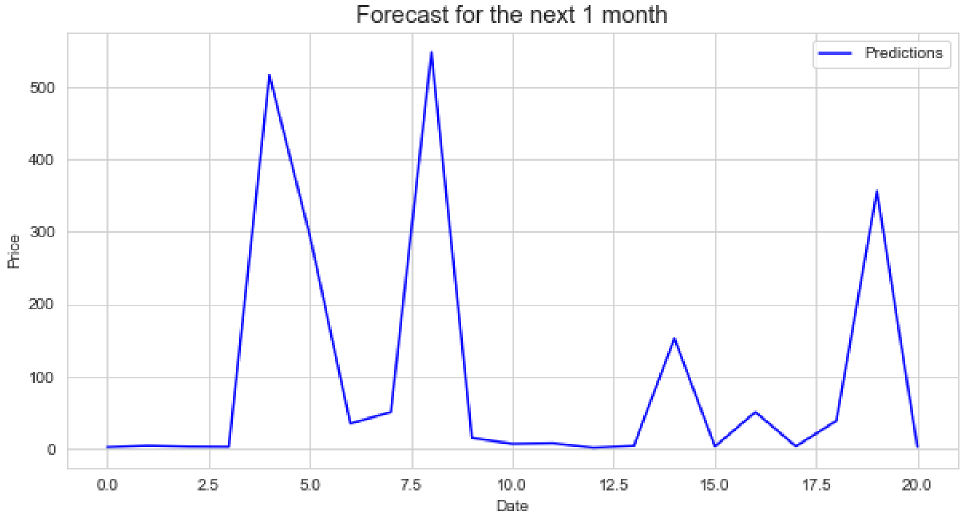
```
Mean Absolute Error: 0.1256
Mean Squared Error: 0.1863
Root Mean Squared Error: 0.4316
(R^2) Score: 1.0
Train Score : 100.00% and Test Score : 100.00% using Random Tree Regressor.
Accuracy: 99.66 %.
MAPE: 0.342409867463272
```

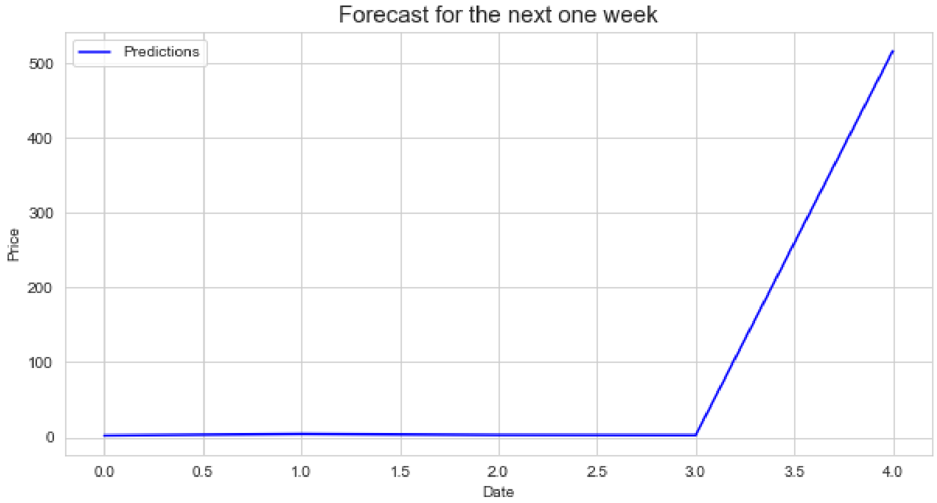
The accuracy of the model is good with very low values for MAE, MSE, MAPE which means that the model is good and can generate close prediction to the actual ones.

PREDICTIONS

Using the model, one year, one month and one-week prediction has been made which can be visualized from the following plots:







CONCLUSION

Using Random Forest, Netflix stock price forecasts have proven to be quite accurate. Across multiple performance metrics, the model performed admirably. It also identified complex data patterns and dealt with nonlinear relationships between variables. The opening price, closing price, highest and lowest prices, and volume are a few of the variables in the dataset that significantly affect the price of Netflix stock. It is possible to obtain reliable results by analysing historical trends, and the Random Forest algorithm is an effective method for predicting Netflix stock prices.