NETFLIX STOCK PREDICTION

Sriya Gorrepati
Computer Science Department
Utah State University
Logan, U.S.
a02370648@usu.edu

Abstract*—The stock market is very unpredictable, any geopolitical change can impact the share trend of stocks in the share market, recently we have seen how covid-19 has impacted the stock prices, which is why on financial data doing a reliable trend analysis is very difficult. This project focuses on forecasting the stocks of Netflix. The dataset chosen contains 5 years of stock information of Netflix i.e, from 5th Feb 2018 to 5th Feb 2022. In this paper we predict the future stock prices of Netflix based on the historical prices using the ARIMA, Random Forest Regression and LSTM prediction models. We analyze the data and make some visualizations on them and then we clean and split the data, using the prediction models we make successful predictions using the existing data. Then measure their accuracy and use evaluation metrics. Finally, we conclude which is the best model.

***Keywords*—ARIMA, Random Forest Regression, LSTM.**

I. INTRODUCTION

The stock market is very unpredictable and fluctuating. Due to this forecasting is considered the most challenging task for the analysts. The process of analyzing data using statistics and modeling to make predictions and inform statistic-decision making is called forecasting. Forecasting has a range of applications in various industries. It has tons of practical applications including weather forecasting, climate forecasting, economic forecasting, healthcare forecasting engineering forecasting, finance forecasting, retail forecasting, business forecasting, environmental studies forecasting, social studies forecasting, and more. Basically, anyone who has consistent historical data can analyze that data with time series analysis methods and then model, forecasting, and predict.

In the past two years Covid-19 has severely affected the stock markets globally, which in turn, created a great problem for the investors. There are several factors that can affect the stock market including recession, geopolitical changes, changes in government regulations, etc., and with the advancement in technology, many organizations and individuals have started using Machine Learning to predict stock prices using historical data to make better decisions in context to investments. Our goal in this paper is to efficiently forecast the stock prices of the Netflix dataset using prediction models. The dataset contains prices of 5 years i.e., from 2018 to 2022. The prediction models which are used in this study are ARIMA[2], Random Forest Regression[1] and LSTM[4] models. All the 3 models are known to forecast the values based on the existing values effectively. Thus, predictions are obtained using these 3 models and we evaluate the best model based on it's evaluation metrics and accuracy measures.

II. DATA SET

The data set is on the Netflix stock price predictions and chosen from the Kaggle. It contains the stock prices of Netflix from the year 2018, February 5th to 2022, February 5th. The goal is to forecast the stocks using the existing prices. The data set values based on a day. It has different attributes and consists of7 columns and 1009 rows. Therefore, it is a huge data set making the task of prediction challenging.

Netflix is a widely used OTT all over the world. During the Covid and lock down people are drawn more towards these OTT platforms. It has been an object of interest to know the stock values of Netflix before and after the Covid. Thus, it is the motivation behind choosing this data set.[6]

III. METHODOLOGY

For this project, I used Google Colab to perform my task and quantitative methodology has been adapted. The following steps have been executed.

A. Importing the required libraries:

The required libraries will be imported such as pandas, seaborn, etc. which will be used throughout the experimental analysis.

Importing the dataset: Once the required libraries are imported, the data will be loaded into the runtime environment.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2018-02-05	262.000000	267.899994	250.029999	254.259995	254.259995	11896100
1	2018-02-06	247.699997	266.700012	245.000000	265.720001	265.720001	12595800
2	2018-02-07	266.579987	272.450012	264.329987	264.559998	264.559998	8981500
3	2018-02-08	267.079987	267.619995	250.000000	250.100006	250.100006	9306700
4	2018-02-09	253.850006	255.800003	236.110001	249.470001	249.470001	16906900
...	...	...	...	...	...	...	...
1004	2022-01-31	401.970001	427.700012	398.200012	427.140015	427.140015	20047500
1005	2022-02-01	432.959991	458.480011	425.540009	457.130005	457.130005	22542300
1006	2022-02-02	448.250000	451.980011	426.480011	429.480011	429.480011	14346000
1007	2022-02-03	421.440002	429.260010	404.279999	405.600006	405.600006	9905200
1008	2022-02-04	407.309998	412.769989	396.640015	410.170013	410.170013	7782400

1009 rows × 7 columns

Picture. 1: Data table

B. Data Cleaning:

The dataset will be cleaned which means checking the null and duplicate values and handling them to perform the data analysis process in an efficient manner. No null or duplicate values are found while checking for the inconsistencies.

C. Model Building:

Three different models will be built namely ARIMA, LSTM and Random Forest Regressor to predict the stock prices of the Netflix stock prices. All three models will be trained using the training subset of the data and will be later validated using the testing subset of the data. The training and testing are divided into 70:30. 70% of the data is used to train the data and 30% is used to test. The target variable will be Closing price as it acts as the base for the last recorded price on a specific date. Since only one target variable has been selected, the model comes under the univariate time series model.

So, this will be considered as a target variable (dependent variable) and other variables will be independent variables. Once all the models are trained and tested, evaluation will be made based on the accuracy score and evaluation metrics. We will see it in more detail in the 'model implementation' section

IV. Exploratory Data Analysis

This section of the report analyzes the data set and to obtain more information and get a clear understanding of the values.

Pandas has been imported for performing data manipulation operations and NumPy has been imported to perform numerical operations on the dataset. There are 7 columns in the dataset, and these are date (shows the date on which the price was recorded), open (opening price of the stock), high (highest price touched by the stock), low (lowest price of the stock on the specific date), close (closing price of the stock), adj close (adjustments made for the stock), volume (volume of the stocks traded on the specific date). The data frame is of the dimension 1009 rows and 7 columns.

Before model building, Exploratory Data Analysis (EDA) is performed on the dataset to understand the data and its statistical characteristics.

	Open	High	Low	Close	Adj Close	Volume
count	1009.000000	1009.000000	1009.000000	1009.000000	1009.000000	1.009000e+03
mean	419.059673	425.320703	412.374044	419.000733	419.000733	7.570685e+06
std	108.537532	109.262960	107.555867	108.289999	108.289999	5.465535e+06
min	233.919998	250.649994	231.229996	233.880005	233.880005	1.144000e+06
25%	331.489990	336.299988	326.000000	331.619995	331.619995	4.091900e+06
50%	377.769989	383.010010	370.880005	378.670013	378.670013	5.934500e+06
75%	509.130005	515.630005	502.529999	509.079987	509.079987	9.322400e+06
max	692.349976	700.989990	686.090027	691.690002	691.690002	5.890430e+07

Picture. 2: statistics

From the statistical characteristics of the dataset, the mean, count, standard deviation, minimum and maximum, and quartile ranges can be estimated which will help in understanding the distribution of the data also.

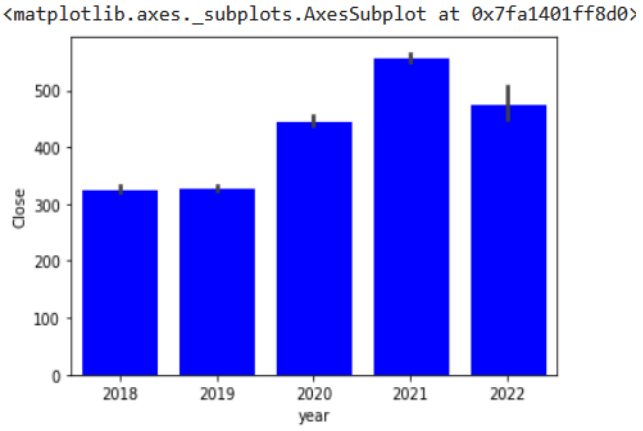
We dropped the "Adj Close" column as it is not relevant while performing the analysis and only other features will be used to make predictions.

To make predictions, there is a need to convert the date (object format) into numerical format by splitting the date format into different columns. After splitting the date column values, this is how the dataset looks like below

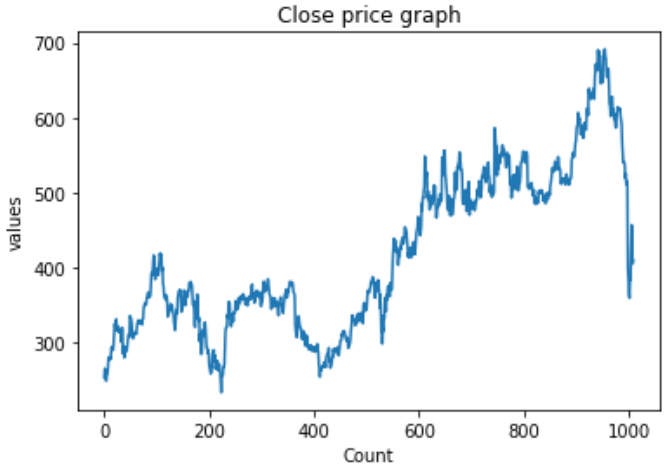
	Date	Open	High	Low	Close	Volume	year	month	day
0	2018-02-05	262.000000	267.899994	250.029999	254.259995	11896100	2018	2	5

Picture. 3

Below graph shows the average closing price with respect to each year. Since 2018, the closing price has been constantly increasing a slight decrease is seen in the year 2022. The graph shows the average values of close attributes with respect to every year.

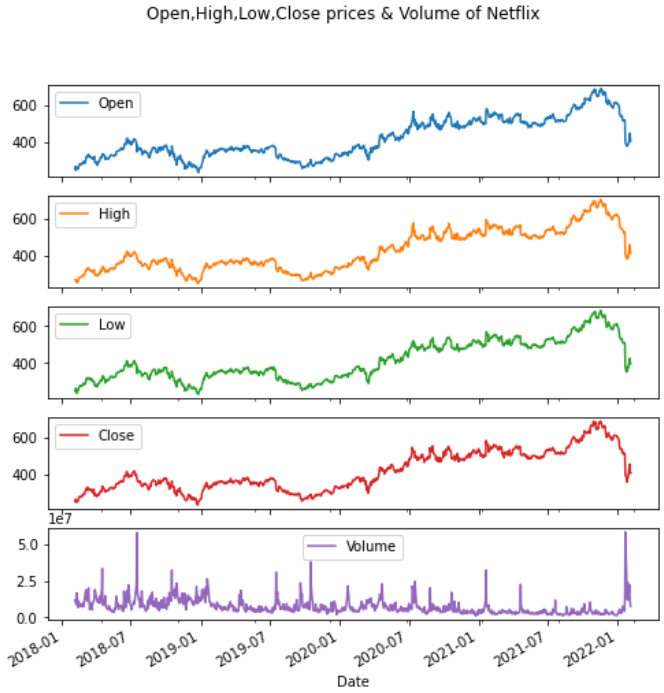


Graph. 1: Bar Graph of close attribute



Graph. 2: Line plot graph of close price

The line plot graph plots the close price values of all the 1000 stocks selected. We can see the increasing plot with fluctuations.



Graph. 3: Price ranges of all the attributes w.r.t years

In the above graph stock price values according to the years. The years are divided into two halves. January to July and again from July to next January. According to my analysis there are 6-month periods for Netflix stocks and they start and end at the period of time.

V. MODEL IMPLEMENTATION

A. Random Forest Model

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. It is used to solve a variety of business problems where the company needs to predict a continuous value: Predict future prices/costs.

Once the data preprocessing was done, the random forest model is implemented. The model is trained on the training data and has been tested using the test data. After implementing the model, the mean square error and accuracy score is calculated as shown below

```
Accuracy Score:  99.17
Mean Squared Error:  21.24
```

Picture. 4: Evaluation metrics of forest regression

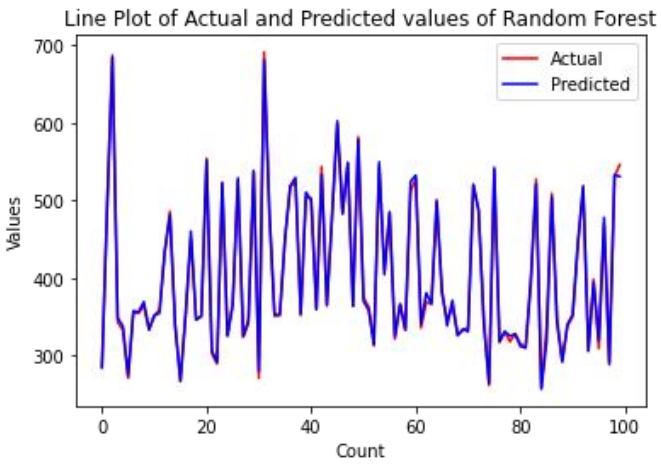
As the accuracy achieved by the Random Forest model is approximately 99% which means that the model will be able to predict the stock price of Netflix with 99% accuracy. A table has been created that shows the Actual vs Predicted values and the actual value in the dataset and the predicted value by the model is almost the same. There is a slight difference only which means that the model is extremely accurate when it comes to predicting the stock prices of Netflix.

The below table is the sample output of predicted and actual values in the random forest regression model. We can see that the predicted and actual values are approximately equal.

	Actual	Predicted
0	286.600006	288.915798
1	366.230011	360.025003
2	325.220001	320.745898
3	351.140015	351.234901
4	286.809998	286.860100
...	...	...
95	521.659973	523.822704
96	550.789978	548.192910
97	357.320007	358.146801
98	346.459991	352.106195
99	497.980011	500.567597

100 rows × 2 columns

Picture. 5: Table showing the actual and predicted values for Random Forest



Graph.3: Line plot of Predicted and actual in Random Forest

The above line plot shows the predicted and actual values of the first 100 values of data using Random Forest model. The red line indicates the actual values, and the blue line indicates the predicted values of the obtained output. The x-plot is the number of stocks considered, and the y-plot is the value with respect to the number of values taken. It can be clearly observed that both the predicted and actual values are on the same line, with this we can conclude that the predictions of random forest are accurate.

B. LSTM (Long Short Term Memory) Artificial Neural Network

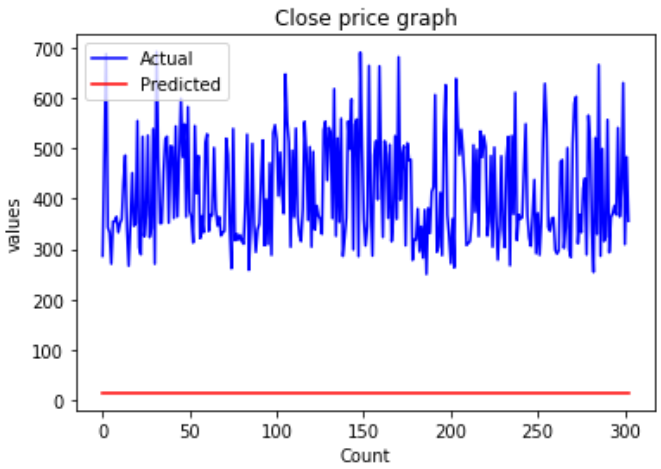
Another model has been implemented named LSTM which is an artificial neural network and is used to solve prediction tasks. LSTM methodology, while introduced in the late 90's, has only recently become a viable and powerful forecasting technique. Classical forecasting methods like ARIMA and HWES are still popular and powerful, but they lack the overall generalizability that memory-based models like LSTM offer. This works based on a sequential model and different layers are added along with an optimizer to make predictions.

The Adam optimizer has been used in this model. After implementing the model, the RMSE of test data, test data Mean Absolute Error of Test data is derived. From the output generated, it is found that the Mean Square Error, Root Mean Square Error, and Mean Absolute Error of the model is high which means that the model is most likely to make mistakes while predicting the values.

```
Test data RMSE:  417.63449963893305
Test data MSE:  174418.575288662
Test data MAE:  403.66140655085457
```

Picture. 6: Evaluation metrics of LSTM

The above picture. 6 shows the evaluation metrics of the LSTM. We can see here that the RMSE, MSE and MAE values are very high. Such high values are not ideal, and it indicates that the LSTM is not proper and failed to make accurate predictions.



Graph. 4. Line plot of Predicted and actual in LSTM

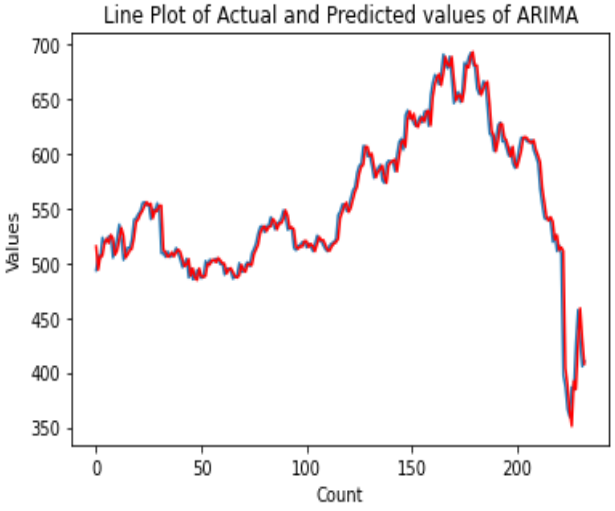
The above graph shows the actual and predicted values of the first 300 values of the data. The x-axis "count" indicates the number of stocks, and the y-axis "values" indicates the value with respect to the number. The blue line is the actual values whereas the red line is the predicted values. We can clearly observe that there is lot of difference between the blue and red

lines, and they are not equal. There are no predictions present in the graph apart from a horizontal line. The horizontal line indicates that the chosen model fits the data very poorly. Thus, LSTM is not working for this forecasting.

C. ARIMA (Auto Rregressive Integrated Moving Average) Model

ARIMA model has also been implemented to forecast the Netflix Stock Prices. ARIMA model is one of the most widely used models and belongs to a class of models that 'explains' a given time series based on its own past values. In simpler terms its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

For the first 200 values, a graph has been plotted that shows the actual vs predicted values where the blue line represents the actual value of the dataset and red line represents the predicted value by the model. Actual and predicted values are almost similar and overlap each other which means that the model is efficiently able to predict the NETFLIX stock prices.



Graph. 5. Line plot of Predicted and actual in ARIMA

In the above graph, x-label indicates the count of the stocks taken and the y-label indicates the values of the stocks. The "red" line indicates the predicted values where as the "blue" line indicates the actual values. It seems in the plot that both the actual and predicted values are on the same line.

	Prediction	Actual
776	514.976684	493.329987
777	493.828858	506.440002
778	506.342790	504.540009
779	506.457955	523.059998
780	518.338816	518.020020
781	522.087621	520.250000
782	518.275548	524.030029
783	525.450472	524.440002
784	522.305377	504.790009
785	507.590978	512.179993

Picture. 6 Table showing actual and predicted values of ARIMA

From the above table, we can see that the actual and predicted values of ARIMA are not precisely calculated by the model. When looked at the first four rows for example, we can see that the actual value is give in the next row previous row's prediction value and the next row prediction value contains the Previous row's actual value. Even though, the graph contains the similar line plot, we can find out that ARIMA model is not performing efficiently from the predicted values and from the evaluation metrics.

```
Mean Absolute Error (MAE) -:  8.18045343850678
Mean Absolute percentage Error (MAPE) -:  0.015458335330422983
Mean square error (MSE) -:  170.6124091735993
Root mean squared error (RMSE): 13.062
```

Picture. 7: Evaluation metrics of ARIMA

From the above evaluation metrics of ARIMA, it can be observed that the MAE, MSE and RMSE of the model are so high living apart MAPE. Only if the RMSE is between 0.2 and 0.5 the model could predict the data accurately. From this we can conclude from this that the ARIMA model is not predicting accurate values and the model is also a failure.

V1. RESULTS

From all the models in the model implementation section we can see that, the random forest regression is giving 99% accuracy and an MSE value of 21.24. This indicates that the model is working and predicting values accurately even though the MSE is a little bit higher. The LSTM model is fitting the data very poorly that all its evaluation metrics are high values. It's RMSE is 417, which indicates that the model is trained so poorly. In the ARIMA model, the values are predicted accurately but still the model is not trained in a proper way. Also, its RMSE is very high which is around 13%.

VII. CONCLUSION

To predict the stock prices of Netflix data set we chose 3 models namely Random Forest regression, LSTM and ARIMA. We first cleaned the data and looked for any data inconsistencies, and we explore the data by making visualizations and analyzing it. We split the data and performed training and testing on it. Then we implemented all the 3 models on the data. By performing the implementation of all the three models, we conclude that "Random Forest Regressor" works better when it comes to predicting the stock prices of Netflix as it showed high level of accuracy and fit the values well. Whereas the LSTM model and ARIMA models showed a high level of error

REFERENCES

[1] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble Machine Learning*, pp. 157–175, 2012.

[2] J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, "Forecasting of demand using Arima model," *International Journal of Engineering Business Management*, vol. 10, p. 184797901880867, 2018.

[3] K. Kandananond, "A comparison of various forecasting methods for autocorrelated time series," *International Journal of Engineering Business Management*, vol. 4, p. 4, 2012.

[4] L. Sun, "LSTM for stock price prediction," *Medium*, 07-May-2021. [Online]. Available: https://towardsdatascience.com/lstm-for-google-stock-price-prediction-e35f5cc84165. [Accessed: 24-Nov-2022].

[5] S. Bouktif, A. Fiaz, A. Ouni, and M. Serhani, "Optimal Deep Learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches †," *Energies*, vol. 11, no. 7, p. 1636, 2018.

[6] https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction