Netflix Stock Price Prediction

Sanjita Chandan Ballapur
PES1UG20CS380
Department of CSE, PES
sanjita811@gmail.com

Shria Guntunur
PES1UG20CS411
Department of CSE, PES
sguntunur@gmail.com

Shruthi Pai
PES1UG20CS416
Department of CSE, PES
shruthipai1924@gmail.com

***Abstract* — Netflix is one of the biggest OTT platforms today and it encompasses a huge share of the stock market. The primary objective of this research paper is to predict the Netflix stock prices at the start of the day. We have done a thorough comparison of the models that have been implemented in previous papers through our literature survey and a summary of the results obtained by these papers and the limitations in their approach have been mentioned in the appropriate section as well. We plan on implementing an LSTM (Long Short-Term Memory) model to try and accurately predict the stock prices and also comparing this model with others like ARIMA, KNN and Regression Classifiers.**

Keywords— Netflix, Stock prices, LSTM, ARIMA, regression

I. INTRODUCTION

Netflix, Inc. is an American subscription streaming service and a production company which was founded in 1997. Netflix is a combination of two words Net (Internet) and Flix (Flick used as an abbreviation for a movie/film). Netflix started a DVD-by-mail rental service which provided an online catalog of movies. Subscribers chose movies and television shows from the company's website, and the shows were then mailed to them in the form of DVDs, along with prepaid return envelopes, from one of the company's more than 100 distribution locations. The present headquarters of Netflix is located in Los Gatos, California. The company has two CEOs, Ted Sarandos and Reed Hastings. Currently, Netflix has over 11,000 employees. As of June 30, 2022, Netflix had 220.7 million subscribers worldwide. Netflix can be accessed via web browsers or via application software installed on smart TVs, set-top boxes connected to televisions, tablet computers, smartphones and many other devices. The company is ranked 115th on the Fortune 500 and 219th on the Forbes Global 2000. As of February 2022, iIt is the second largest entertainment/media company by market capitalization.
Stocks, or equity, are a type of security that gives stockholders a share of ownership in a company. Stocks are of two types—common and preferred. The difference is while the holder of the former has voting rights that can be exercised in corporate decisions, the later doesn't. Investors buy either kind of stocks for capital appreciation, dividend payments as well as a chance to influence the decisions of a company. Stocks offer investors the greatest potential for growth (capital appreciation) over the long haul. However, there's no guarantee that a company does well, so one can lose the money they invest in stocks easily. The risks of stock holdings can be offset in part by investing in a number of different stocks. The netflix stock symbol/ticker is NFLX. Netflix is currently worth around 106.04 billion USD. The annual revenue generated by the company for the year 2021 was 29,697,844 USD.
Time-series data is a sequence of data points collected over time intervals, allowing us to track changes over time for repeated measurements. In the dataset that we have chosen, data has been collected from May 23, 2002 to June 3,2022 which is over a period of 20 years. The features taken include the opening and closing prices for a particular day as well as, the highest and lowest stock price. The last column mentions volume which indicates the total number of shares traded on that day. In this paper, we list our findings from comparisons made for different models that have been implemented in various papers along with our proposed methodology.

II. RELATED WORK

In this section, we summarize and explain the approach followed in various research papers. Furthermore, the final result and any limitations have also been mentioned.

The aim of [1] was to predict the opening, lowest and highest price of the stock on the next day. Three datasets were used, namely, Shanghai composite index,two stocks of PetroChina(on Shanghai stock exchange) and ZTE(on Shenzhen stock exchange). Each dataset was divided into training and test set in the ratio of 4:1. The data was normalized using the min-max method which improved accuracy. The model proposed was an Associate net model as compared to LSTM and DRNN.
Long short-term memory network (LSTM) is a form of RNN. Its structure has three gates-the input gate, forgetting gate and output gate. Information not selected by the model will be forgotten through the forgotten gate. Sigmoid function is the default activation function while a tanh function is used to update the state of neurons. DRNN (LSTM-based deep recurrent neural network) is a variant of RNN. It operates in the same way as LSTM but here, the dropout method was used. Dropout refers to dropping out units (both hidden and visible) in a neural network. Neurons in each layer are randomly deleted to prevent overfitting. Since the three predictions are associated with each other, each of these are predicted by different networks. A structural model of multi-value associated neural network (which was based on the deep recurrent neural network) based on LSTM was designed. The model combines the output of the first branch(opening price) and the second branch(lowest branch) to give as input to the third

branch(highest branch) which finally gives an appropriate output.MSE is used to measure the quality of the model. The Adam optimization algorithm was used in the training phase. Various graphs were plotted to figure out the best step size and as a result, 20 was chosen as the step size. Interesting results were found during the training phase. The LSTM model faced multiple fluctuations during the training process but Associated Net and DRNN were very stable. It was seen that LSTM was overfitting the data so it had the lowest average accuracy. Due to Associated Net being a very complex model, it had the highest average MSE as it needed a large number of iterations after which, the total loss of the model gradually reduced. The general findings was that the more the training data, the better is the model fitting effect.
In the testing phase, MAE(average of the absolute error) was used while the average accuracy was equal to 1-MAE. The result was that DRNN performed better than LSTM but Associated Net fit the curve of the real data better than DRNN. The deviation from the data for Associated Net was smaller than that of DRNN. Finally, Associated Net had a higher average accuracy than the other two models. Not only can it predict multiple values simultaneously, the average accuracy was over 95%. A limitation found was that the loss calculation method used does not take into account the relationship between each sub-loss.

This paper [2] uses regression techniques for prediction of stock price trend by using a transformed data set in an ordinal data format. The original dataset consists of heterogeneous data types used for handling currency values. The transformed data set contains only a standardized ordinal data type which gives a measure of the rankings of the stock price.
Stock price data was collected from companies in Bursa Malaysia for the years 2003-2010. The selected features were Net Tangible Asset (NTA), Liquid Asset (LA), Debt to Equity (DE), Altman Z-Score (ZS) and Asset Turnover (AT). Statistics on the companies and the dataset features were generated through analysis. The data was pre-processed to remove out-of-bound values. Pre-processed data containing real-valued data was standardized into percentage oriented data. A percentage to ordinal conversion table contained the ranges of percentage values associated with their ordinal enumerated values.
Linear Regression, Additive Regression, Regression by Discretization, Simple Linear Regression and SMO Regression classifiers were used as predictive analytic on the dataset. A percentage split was used to divide the data into training data and testing data proportionally. Training data was used to formulate regression rules and that was used on the testing data for making predictions on future stock price trends.
It was observed that the result improved after transforming the data from numerical to ordinal values. All but one classifier used gave lower error rate and higher correlation coefficient for the transformed dataset. The SMO Regression technique performed the best with a correlation coefficient of 0.6079 and with a mean absolute error of 0.5462 and a root mean square error of 0.8164.
In conclusion, this research shows that the outcomes of regression techniques can be improved by standardizing the data into a common data type through a transformation process. The use of an ordinal data type for prediction based on ranking system provides a different dimension for predicting outcomes. SMO Regression technique has outperformed the other regression techniques in the experiment. So, it can be inferred that outcomes are favourable when less structured data are transformed into more structured.

Another paper that we were intrigued by was [3] in which the research used Netflix stocks historical data for the past five from 7 April 2015 to 7 April 2020 to compare the results of auto ARIMA model and two customize ARIMA (p,D,q) models. The data in the dataset describes the date, open which is the price at the beginning of the day, high which is the highest price during the day, low which is the lowest price during the day, close which is the price at the end of the day, adjusted closing which is the price of stock's closing price amended to accurately reflect that stock's value after accounting for Netflix actions, and the volume which is the number of stocks of a security traded during that day. Determining the model accuracy and the comparing between the several experiments in the model will be based on calculate Autocorrelation Functions (ACFs), Partial Autocorrelation Function (PACF) as well as Mean Absolute Percentage Error (MAPE). Auto Regressive Integrated Moving Average (ARIMA) is a model describes time series given based on observed value which can be used to forecast future values. Appling ARIMA models on Any time series show patterns with no random white noise.To generate short-term forecasts, ARIMA models showed efficient capability and outperformed complex structural models. A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where: p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation.
Forecasting three models of ARIMA, Auto ARIMA (4,1,4) and ARIMA (1,2,33) showed the same prediction which predicted that stocks will go up while in ARIMA (1,1,33) has different prediction where predicted that the stocks will remain the same. Comparing the accuracy results by calculating Mean Absolute Percentage Error (MPE) showed no much difference between the three models, where the accuracy of auto ARIMA is 98.88 %, ARIMA (1,1,33) is 99.74% and ARIMA (1,2,33) is 99.75%.
After several tests ARIMA (1,1,33) showed accurate results in its calculating values and it showed continuity in value. So it can be inferred that the potential of using ARIMA model on time series data to get accurate prediction on stocks data will help investors in stocks in their investment decisions.

III.PROBLEM STATEMENT

The dataset used in this project has been taken from Kaggle which is available at the following link: https://www.kaggle.com/datasets/meetnagadia/netflix-stock-price-data-set-20022022

Our problem statement is to evaluate the Netflix stock prices for the past 20 years(2002-2022) and to accurately predict the opening stock price of the day.

A. *Dataset*

The data contains 7 features. It is a dataset for the stock price of Netflix from 2002 to 2022 .It was collected from Yahoo Finance.
The ‘Open’ attribute is the price at which the financial security opens in the market when trading begins.The ‘High’ attribute is the highest price at which a stock traded during a period. The ‘Low’ attribute is the lowest price at which a stock traded during a period. The ‘Close’ attribute is the price that generally refers to the last price at which a stock trades during a regular trading session.The ‘Adj Close’ price amends a stock's closing price to reflect that stock's value after accounting. ‘Volume’ measures the number of shares traded in a stock or contracts traded in futures or options.

B. *Exploratory Data Analysis and Visualizations*

The dataset has 5044 rows and 7 columns in total. Out of the 7 columns, it was found that 5 columns were of type float64, 1 column was of type object which was Date and 1 column was type of int.
After data preprocessing it was found that there were no missing values in the data. After plotting the box plots, it was found that all the columns have a high number of outliers. Since it is a stock price dataset, the outliers are significant and have not been removed. In Fig. 1. , the correlation plot shows that there is strong positive correlation between all the columns except Volume. Volume has a weak negative correlation with the other columns because it is the values of the shares traded and not the price of the stocks.

Fig. 1. Correlation plot

In Fig. 2., it was noticed that the opening and closing stock price was increasing for a few years but in recent times it has started dropping this may be because of the investors losing faith in the stock market and also due to drop in subscriber count.

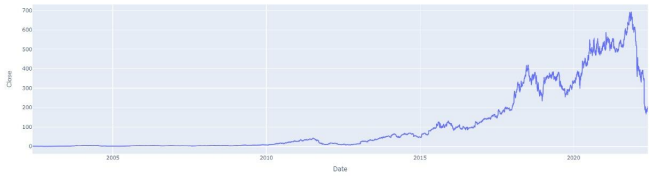


Fig. 2.

In Fig. 3., it was noticed that the shares traded dropped during 2019-20.The onset of the covid-19 pandemic brought the worst financial crisis since the great recession of 2008. The stock market plummeted further putting people at risk of their investments and income. Due to this people lost faith in the stock market and refused to buy stocks which is why we see a continued decrease in these years. Netflix decided to reduce their subscription fees in order to increase the users on their platform which is why we see a sharp rise in 2022.

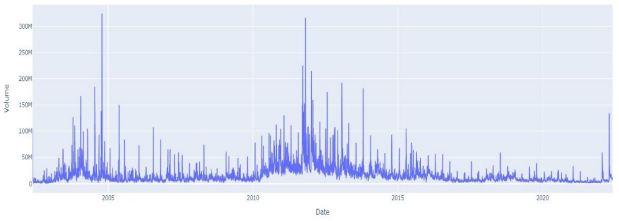
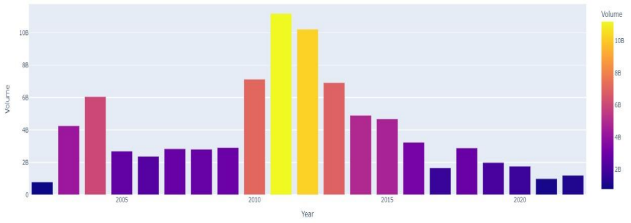


Fig. 3.

In Fig. 4., from the bar graph, we noticed that the maximum number of shares had been traded in the year of 2011. In 2021, there was an all time low of shares traded because of high subscription fees, reduced number of users and numerous competitors in the market as well as the reopening of movie theatres.



\

Fig. 4.

In Fig. 5., the bar plot is shown for the first 30 rows and we can see that the opening price is usually more than the closing price due to the fluctuation between the balance of supply and demand. Also,development of after-hours trading (AHT) has had an effect on the price of the stock between the closing and opening prices of the following day.

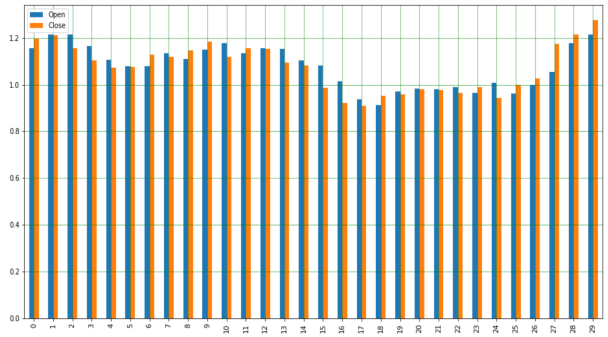
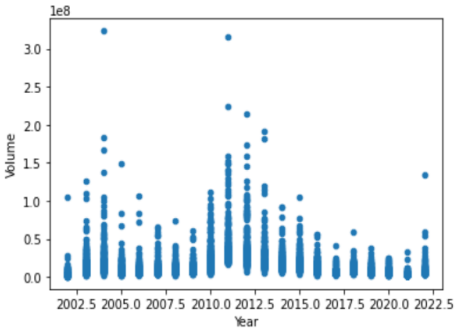


Fig. 5.

In Fig. 6a., it was noticed from the scatter plots that outliers are seen for specific years as this is very probable because this is real-time stock data.



As seen in Fig. 7., pairwise scatter plots were plotted for all attributes.It was seen that, High and Low had an almost linear relationship as well as Close and High. As there are many data points, the graph is clustered.

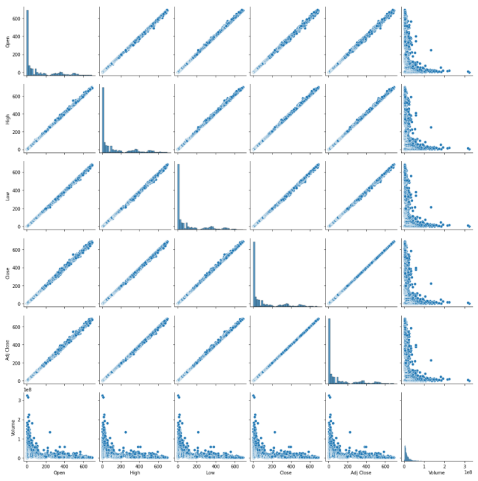


Fig. 7.

IV. PROPOSED METHODOLOGY

After reading various papers and exploring the different models used, we have decided to implement a model after comparing the accuracies of LSTM, ARIMA and regression techniques. As of now, the next step is to work with the LSTM model and see how it works against our chosen dataset.

V. REFERENCES

[1] Guangyu Ding,Liangxi Qin,"Study on the prediction of stock price based on the associated network model of LSTM",International Journal of Machine Learning and Cybernetics (2020)

[2] Han Lock Siew, Md Jan Nordin,"Regression techniques for the prediction of stock price trend",IEEE-2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)

[3] Shakir Khan , Hela Alghulaiakh,"ARIMA Model for Accurate Time Series Stocks Forecasting",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7, 2020