Time Series Analysis of Stock Market Data

Swathi V
Dept. of Engineering Design
Indian Institute of Technology Madras
Chennai, India
ed18b034@smail.iitm.ac.in

***Abstract*—A time series data is one that is observed and recorded over an interval of time at consistent intervals. Hence, the order of the observations is preserved and is important for forecasting the trend of a variable. In this report, we analyse and compare multiple models that can be used for time series analysis, and quantify their performance of financial data obtained using the stock prices of six different companies.**

***Index Terms*—Time series, forecasting, exponential smoothing, ARIMA, regression, financial data, stock prices**

I. Introduction

Time series forecasting has been fascinating to humans for thousands of years, as the ability to predict something from the future was and is considered a great power. A good prediction can seem almost magical, while bad forecasts can be dangerous. For example, the Chairman of DEC said in 1977, three years before IBM produced the first personal computer, "There is no reason anyone would want a computer in their home." Forecasting is obviously a difficult activity, and businesses that do it well have a big advantage over those whose forecasts fail.

The ease of forecasting a particular variable can vary widely depending on several factors such as the amount of data available, and the extent of understanding we have on the contributing factors. Specifically, quantitative forecasting techniques can be applied only when numerical information about the past is available, and it is reasonable to assume that some aspects of the past patterns will continue into the future.

Financial data is a very good example of a time series which can be forecast, as it satisfies both the above mentioned conditions. In this paper, we will perform time series analysis and forecasting on the stock prices of six different companies (Cognizant, HCL Technologies, HDFC Bank, ICICI Bank, Infosys, and SBI), the history of which is available for a period of two years (2019-2021).

This paper will start with a mathematical overview of time series analysis, and the same can be analysed to extract meaningful trends. We then look at various techniques that can be used for forecasting, and the conditions under which each technique can be used accurately. We will then analyse the trends and try to predict stock prices of the six companies mentioned above.

II. Time Series Analysis and Forecasting - A Mathematical Overview

A time series is any data that is observed and recorded over an interval of time at consistent intervals. The main aim of time series analysis is to understand the underlying trends in the data so as to try and predict how it will vary in the future (forecast). Once the data is gathered, the major steps involved in forecasting are:

- Exploratory data analysis (trying to extract information from existing data)
- Choosing and fitting models
- Using and evaluating a model

A. Exploratory Data Analysis

Preliminary analysis of time series data involves checking for a few properties, such as autocorrelation, stationarity, and seasonality. These tests help us determine the best methods for forecasting data.

1) Stationarity: Stationarity is the property of exhibiting constant statistical properties (mean, variance, etc.). For example, if the rolling mean of a time series is increasing or decreasing (or varying in any fashion) over time, then the series is not stationary. In order to find if a given time series is stationary, we use the *augmented Dickey-Fuller unit root test*, the null hypothesis of which states that the time series is not stationary.

2) Seasonality: A seasonal pattern occurs when the time series is affected by factors such as the time of the year or the day of the week. If a similar patter is observed every season, then the time series is said to be seasonal.

3) Autocorrelation: Autocorrelation measures the linear relationship between lagged values of a time series. The autocorrelation coefficient for a lag k is calculated as:

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})} \tag{1}$$

where T is the length of the time series. When the data has an observable trend, the autocorrelation for small lags tend to be high and positive, and slowly decrease as the lag increases.

4) Detrending data: Detrending is used to remove the underlying trend in the data. It is done by subtracting the rolling mean and diving by the rolling standard deviation from the time series. The detrended data can be subtracted from the raw data to obtain the underlying trend without the added noise.

B. Models for time series analysis

Various models can be used for time series analysis depending on the properties of the series. The three major techniques

that will be analysed in this paper are - exponential smoothing techniques, ARIMA models, and regression models.

1) Exponential Smoothing Techniques: Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations get older. The more recent the observations, the higher their weights.

Simple exponential smoothing: This method is suitable for forecasting data with no observable trend. The predicted value for a time $t+1$ is given by:

$$y_{t+1} = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j y_{t-j} \tag{2}$$

Holt's linear trend method: Holt extended simple exponential averaging to produce a linear forecast which is given by:

$$y_{t+h} = l_t + hb_t \tag{3}$$

$$\text{where the level } l_t = \alpha(y_t) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{4}$$

$$\text{and the trend } b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{5}$$

This forecast function is no longer flat but trending.

Damped trend method: The problem with Holt's Linear trend method is that the trend is constant in the future, increasing or decreasing indefinitely. The damped trend method adds a dampening parameter, so that the trend converges to a constant value in the future. The forecast values are given by:

$$y_{t+h} = l_t + (\phi + \phi^2 + \cdots + \phi^h)b_t \tag{6}$$

$$\text{where } l_t =_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \tag{7}$$

$$\text{and } b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1} \tag{8}$$

where ϕ is the damping parameter.

2) ARIMA Models: Auto Regressive Integrated Moving Average models are a class of models used to predict a given time series based on its own past values. An ARIMA model is characterised by three terms: p (the order of the AR term), q (the order of the MA term), and d (the number of differencing required to make the series stationary).

The d term: The first step towards building an ARIMA model is to make the series stationary, by differencing it. d is the minimum number of differencing required to make the series stationary.

The p term: p is the order of the auto-regressive term. It refers to the number of lags to be used to predict the subsequent values.

The q term: q is the number of lagged forecase errors that should go into building an ARIMA model.

The ARIMA model with coefficients p, d, q is given by:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{9}$$

where y' is the differenced time series.

3) Regression Models: The exponential smoothing and ARIMA models are specific to time series. In addition to these, normal regression models (especially non-linear regression models such as SVRs) can be used to predict the time series. In this case, for every output y_t of the time series, the values $y_{t-1}, y_{t-2}, \cdots, y_{t-p}$ are taken as the feature variables, where p is a tunable parameter. In this analysis, regression will be tried using SVRs, using three previous data points as features.

III. Data

The data used for this analysis is the stock market data, which contains the *opening*, *closing*, *high* and *low* prices of the stocks of six companies over a span of two years. The six companies are: Cognizant, HCL Technologies, HDFC Bank, ICICI Bank, Infosys, and SBI.

Except the stock prices of Cognizant, all other prices are given in INR. Hence, the USD to INR conversion rates were used to convert the stock prices of Cognizant to INR.

IV. The Problem

For each company, we first analyse the data, and then try to predict the stock prices for the next ten days using the selected method. To select the best method for the same, we compare the different methods mentioned above. The steps followed for the same are:

- Data cleaning
- Data analysis and visualization
- Fitting multiple models and comparing their performance
- Using the selected model to make predictions

A. Data Cleaning

Some columns contain missing values in the given dataset, which are interpolated using *linear interpolation*, since it is a time series data (ordered).

B. Data Visualization and Analysis

Complete data visualization: The stock prices of each company are plotted using a candlestick plot, where the volume of stocks traded is also shown in a bar graph for each date. In these plots, the candlestick is coloured *red* or *green* depending on whether the closing price is below or above the opening price. The plots also contain the rolling average plotting over a 15 day window, the visualization of which can reveal the underlying trend in the data without corruption by noise. These plots are shown in Fig. 1 to Fig. 6 for the six companies respectively. These plots show that none of the six stock prices are stationary, as the rolling mean is not constant. In general, the stock prices have all increased during the period of observation.

Correlation between various fields: We analyse the correlation between the various prices (*open*, *close*, *high*, *low*) for each company. The correlation between these values is shown in Fig. 7 for the stocks of Cognizant. We see that all four price fields are very closely related, which makes it sufficient to predict just one of the four. In this case, we will perform all further analysis using the *closing* price. Similar trends were observed for stocks of all other companies.

Checking for seasonality: To check for seasonality in the Cognizant stock prices, we plot the stock prices (closing) varying within a year. This plot is shown in Fig. 8. We see that

Fig. 1. Visualization of the stock prices of Cognizant

Fig. 2. Visualization of the stock prices of HCL Technology

Fig. 3. Visualization of the stock prices of HDFC Bank

Fig. 4. Visualization of the stock prices of ICICI Bank

Fig. 5. Visualization of the stock prices of Infosys

Fig. 6. Visualization of the stock prices of SBI Bank

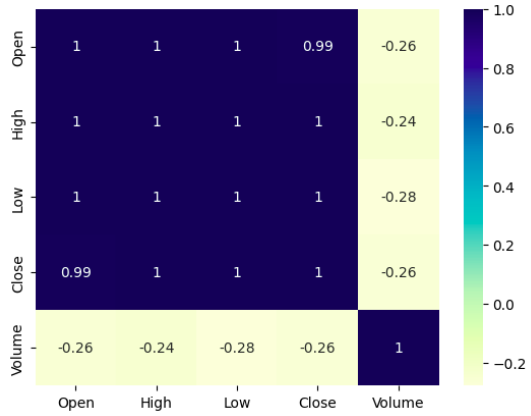

Fig. 7. Correlation between various fields in the Cognizant stock dataset

the time series does not have a seasonal dependence, which is expected.

Checking for autocorrelation: The stock prices of each company are checked for autocorrelation, the result of which is shown in Fig. 9 for Cognizant. Stationarity can also be concluded from the autocorrelation plot, wherein a stationary time series has very high positive values of autocorrelation for smaller lags, which quickly declines to 0. This is not observed in Fig. 9. The blue band shows the 95% confidence interval for the values.

Comparing the stock prices of all companies: To compare the stock prices of the six different companies, they were plotted in a single graph, shown in Fig, 10. We observe that there is a huge variation in the stock prices of cognizant, which is not the case with the other companies. This might be a result of the larger price scale, as well as the fluctuations in the USD to INR conversion rate.

In order to check for relationships between the stock prices of the different companies, we plot the correlation heatmap, shown in Fig. 11. We see that there is a strong correlation between stock prices of HCL Technology, HDFC Bank, ICICI Bank, Infosys, and SBI. However, this does not affect the time series prediction, since we will be predicting the stock prices of each company separately.

Checking for stationarity of the data: To check for stationarity of the data, we can use, in addition to the visualizations in Fig. 1 and Fig. 9, the plot of the underlying trend in the data. This is obtained by first detrending the data, which is then subtracted from the raw data. The plots obtained are shown in Fig. 12, which clearly shows that the data is not stationary.

C. Fitting multiple models and comparing their performance

To fit each model, we divide the data into two parts. The first part of the data (all except last ten observations) will be used as the training data, while the last ten observations will be used as the test data. The model which gives the best performance on the test data (lowest mean absolute error) will be used to predict the stock prices for the next ten days. For the cognizant data, these periods are:

- Training: 02-01-2019 to 20-09-2021
- Validation: 21-09-2019 to 04-10-2021
- Test: 05-10-2021 to 14-10-2021

To determine the best model to forecast the given time series, different models were fitted and their mean absolute errors measured, which include - simple exponential smoothing, holt's linear trend, damped trend, ARIMA, and regression using SVR. The selection of model parameters for the Cognizant stock data is explained below.

1) Exponential Smoothing Methods: For all exponential smoothing methods, the parameters were not given manually, hence the parameters which best fit the given data were chosen automatically.

2) ARIMA Models: To determine the value of d for the ARIMA model, augmented Dickey-Fuller unit root test was used. The p-values were analysed, and d was chosen to be 1 for the given time series. To find the p and q values, the partial autocorrelation and autocorrelation were plotted for the first order differenced time series. This is shown in Fig. 13, from which both p and q were chosen to be 1. Hence, a (1,1,1) ARIMA model was chosen to fit the data.

3) Regression using SVR: To predict every output value, three previous values were taken as input features, and a linear SVR was trained.

The results of all the above said methods along with the ground truth is shown in Fig. 14 to Fig. 19 for the stock prices of all six companies respectively. The errors for each of these models are mentioned in the legend.

D. Using the best model to make predictions

In all six cases, SVR performed better than the other models in terms of mean absolute error. Hence, SVR models are trained with the complete dataset and used to predict the stock prices for the next ten days, for each of the six companies. The results of these are shown in Fig. 20 to Fig. 25.

The stock prices of Cognizant, HCL Technology, HDFC Bank, ICICI Bank, Infosys, and SBI were predicted with a mean absolute error of approximately INR 48, 15, 12, 6, 16, and 5 respectively. The mean absolute percentage errors were 0.8%, 1.2%, 0.7%, 0.9%, 0.9%, and 1.2% respectively.

The results are all linear predictions for the next ten days, which is because of the fact that the model uses the predicted values and not the actual values as input for days 2 and above. However, in a real life scenario, true data will be fed into the model for every subsequent day, which will lead to non-linear predictions.

V. Conclusions

In this report, various methods for time series forecasting were analysed, and their performances were compared on financial data. Support vector regression was used to accurately forecast the market prices, using the prices of three previous days.

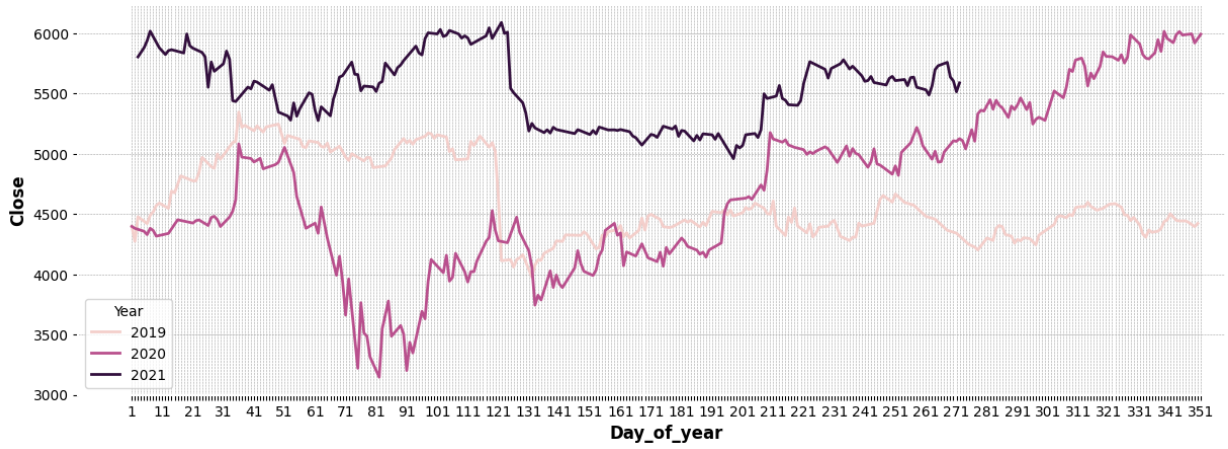

Fig. 8. Seasonal plot of the cognizant stock data

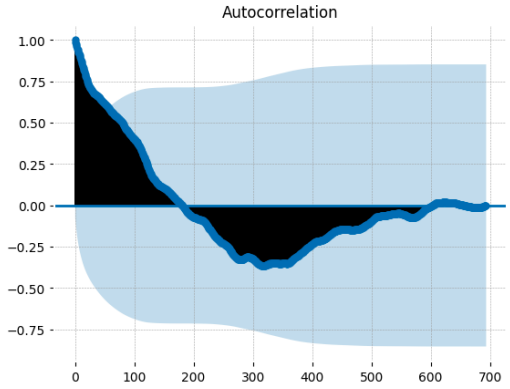

Fig. 9. Autocorrelation for the cognizant time series, plotted for various lag values

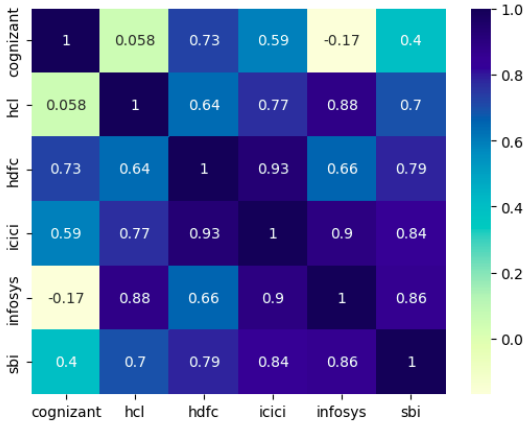

Fig. 11. Correlation between stock prices of all companies

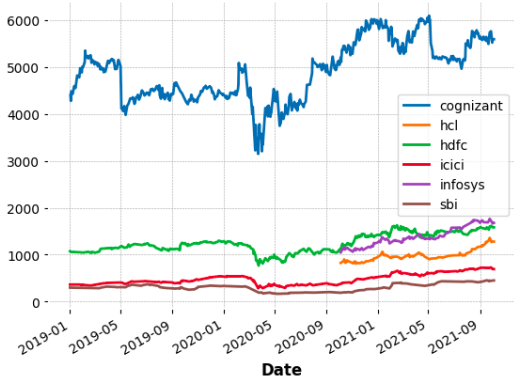

Fig. 10. Comparison between stock prices of all companies

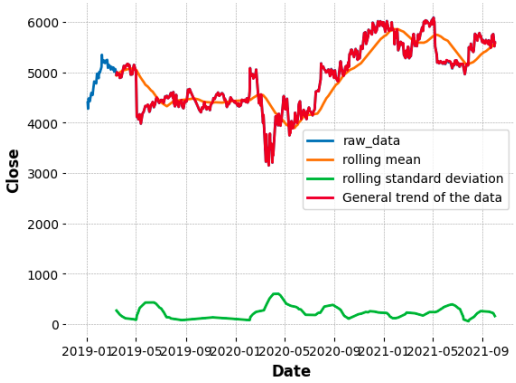

Fig. 12. Checking for stationarity in the cognizant dataset

A. *Avenues for Future Work*

Other regression methods, such as linear and polynomial regression can be used to predict stock prices. In addition, instead of using data from 3 days as input, the number can be tuned as a hyperparameter.

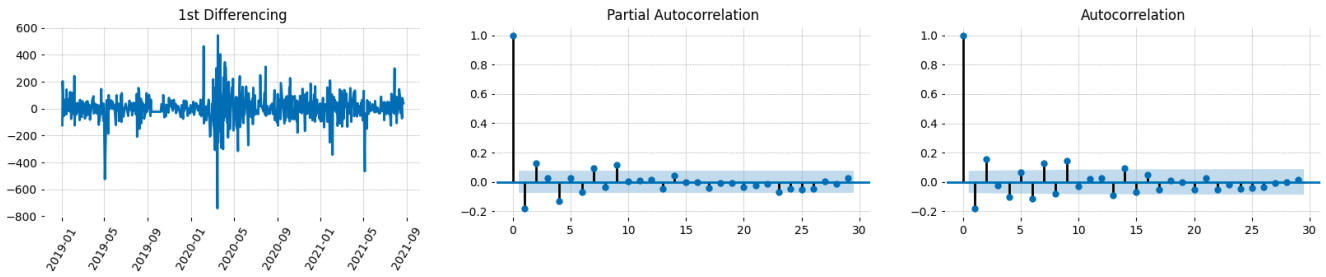

Fig. 13. Partial autocorrelation and autocorrelation for the once differenced time series

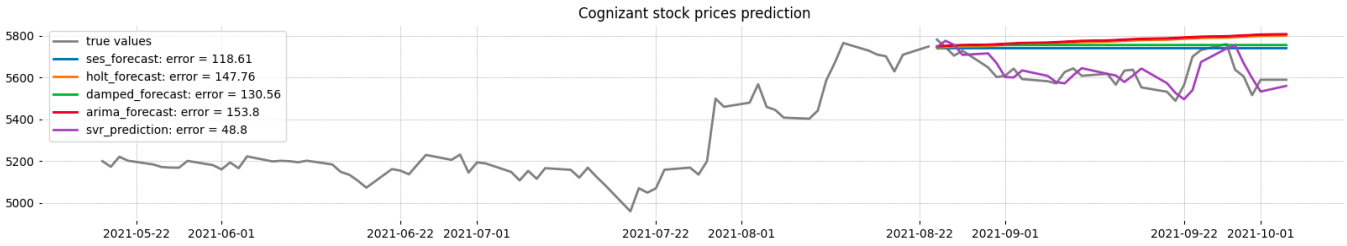

Fig. 14. Results of various models fit on the Cognizant stock data

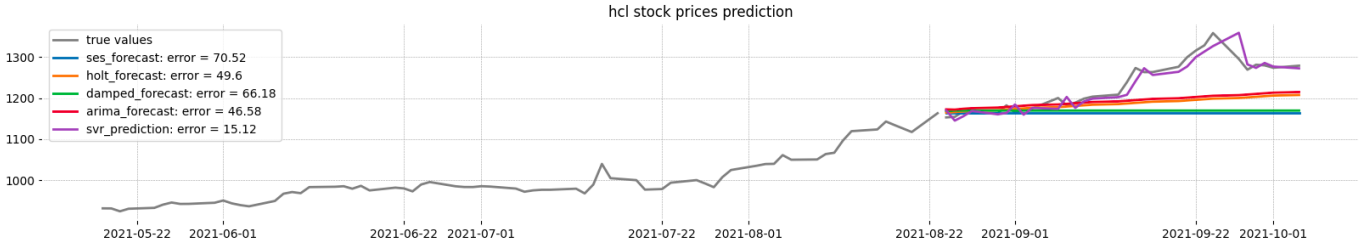

Fig. 15. Results of various models fit on the HCL stock data

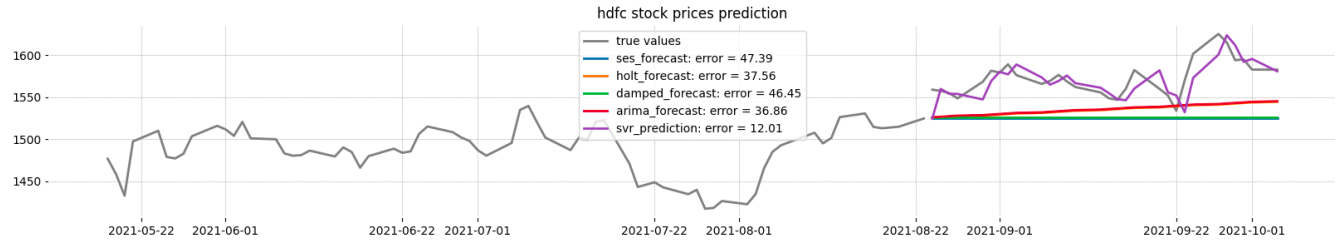

Fig. 16. Results of various models fit on the HDFC Bank stock data

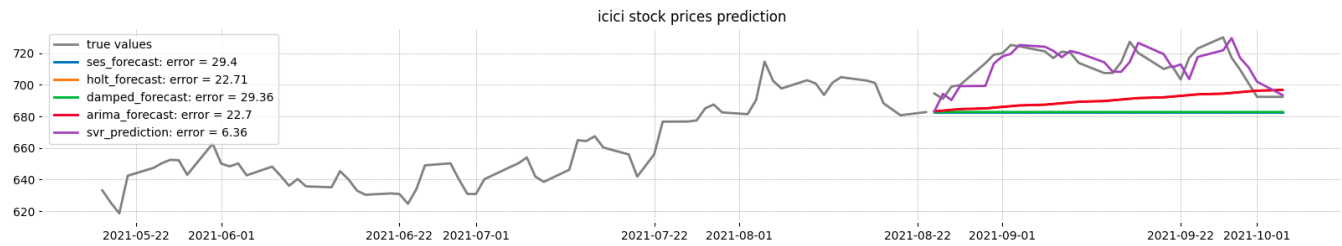

Fig. 17. Results of various models fit on the ICICI Bank stock data

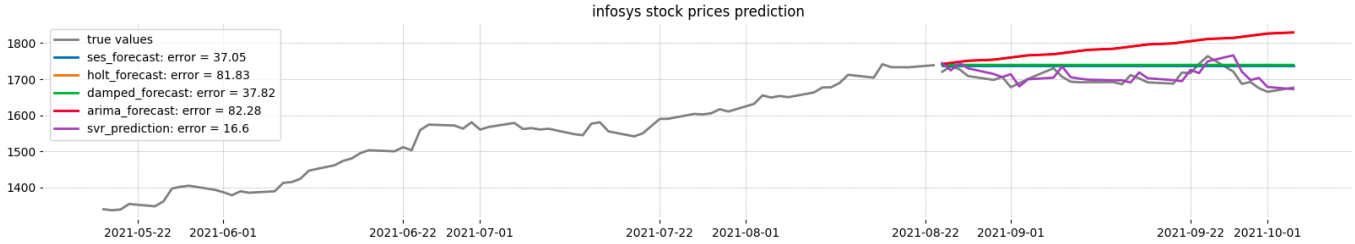

Fig. 18. Results of various models fit on the Infosys stock data

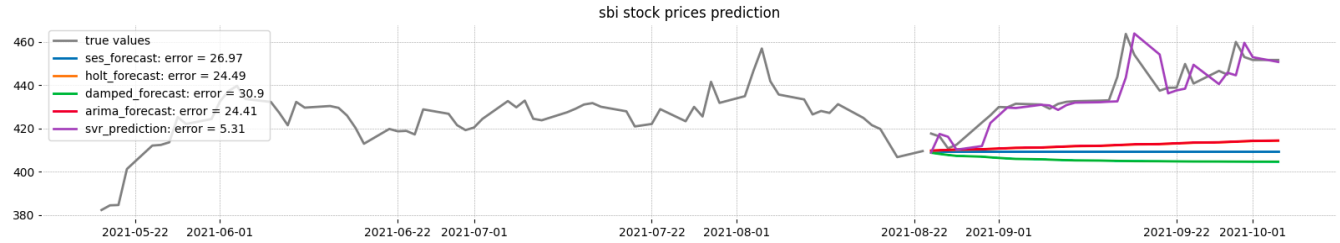

Fig. 19. Results of various models fit on the SBI Bank stock data

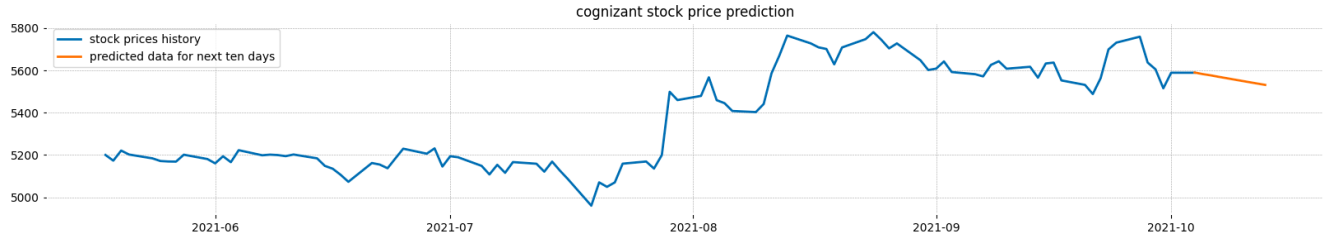

Fig. 20. Stock price prediction for Cognizant

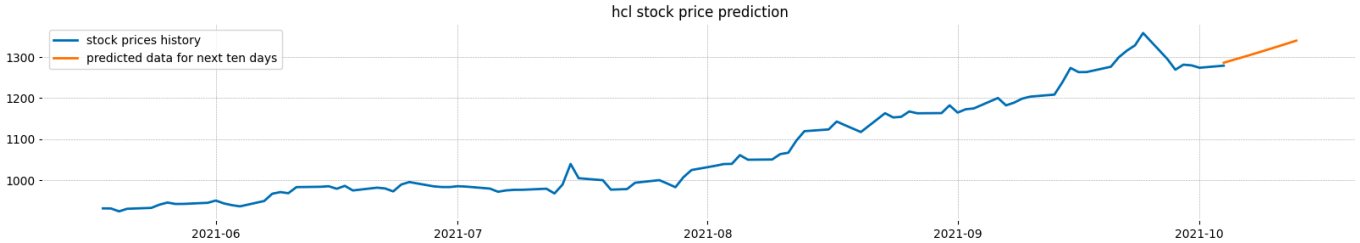

Fig. 21. Stock price prediction for HCL

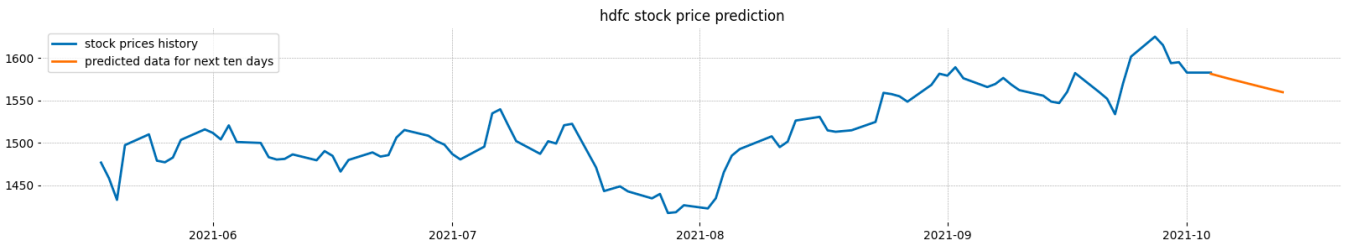

Fig. 22. Stock price prediction for HDFC Bank

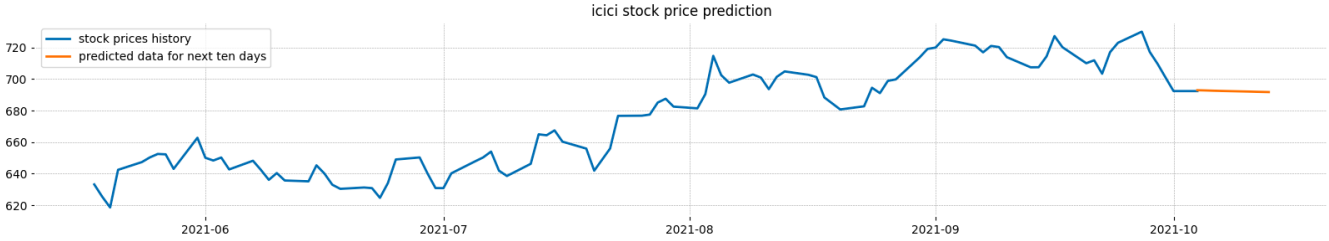

Fig. 23. Stock price prediction for ICICI Bank

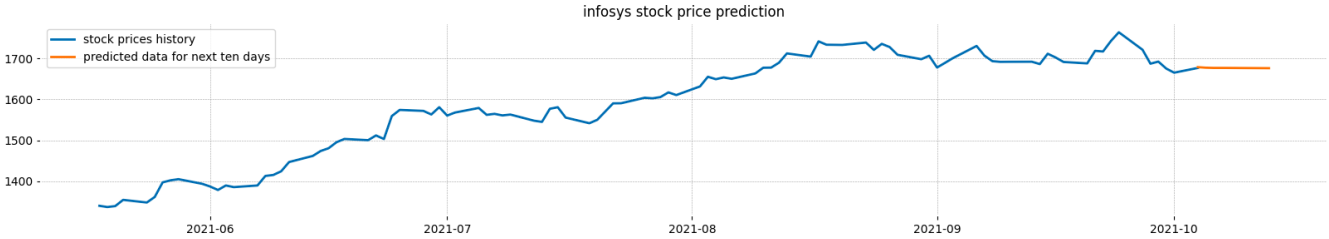

Fig. 24. Stock price prediction for Infosys

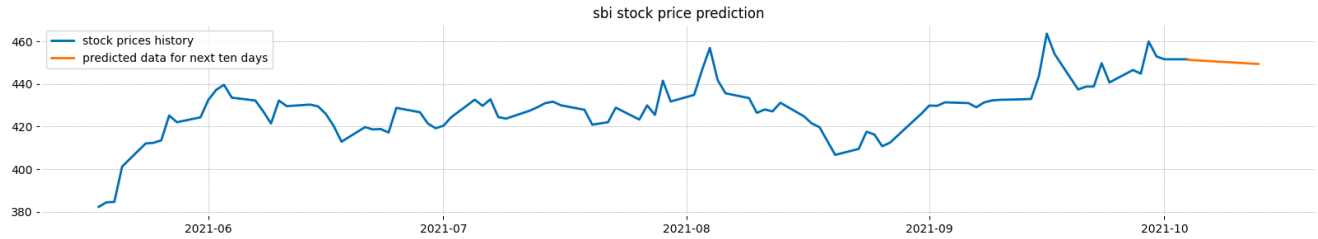

Fig. 25. Stock price prediction for SBI Bank

REFERENCES

[1] Rob J Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice*, Monash University, Australia.

[2] Selva Prabhakaran, *ARIMA Model – Complete Guide to Time Series Forecasting in Python*.