# ARIMA for Time Series Analysis

Vaibhav Mahapatra
*Data Science Department*
*Indian Institute of Technology, Madras*
Chennai, India
me19b197@smail.iitm.ac.in

***Abstract*—The AutoRegressive Integrated Moving Average (ARIMA) model is a famous and widely used forecasting method for time-series prediction. ARIMA models are capable of capturing a suite of different standard temporal structures in time-series data. In this article, we will be performing extensive data analysis on stock prices of 6 organizations over a span of 3 years (2019-2021). In the IT sphere, we will be focus on the stock price fluctuation of 3 firms, namely Cognizant, HCL Technologies and Infosys. In the banking sphere, we'll look at 3 institutions, namely HDFC bank, ICICI bank and SBI bank. We will also be analysing the fluctuation in the value of the INR currency when compared to 1 USD over the span of 2 years.**

**Further, we'll use Random Forest Regressor and Moving Average based Techniques like ARIMA to build predictive models to estimate quantities like volume of shares traded in a day, share price at the end of the day, value of 1 USD in INR on a certain day, etc.**

***Index Terms*—Time series analysis, Regression, moving average, kurtosis, features, predictors, kurtosis, mean squared error.**

## I. Introduction

In today's world financial instruments play a significant role in determining the economic health of a country. The rate at which these instruments are traded between certified entities directly affects the rate of economic growth. They are also an important means to generate capital for corporate and public organisations alike. With high volumes of these instruments being traded on a daily and hourly basis, performing extensive data analysis plays a crucial role in trading these instruments in a manner to minimise risk and maximise returns. Trading of these instruments is done on a platform which can be referred as the stock market. In its broadest sense, the stock market refers to the collection of exchanges and other venues where shares of publicly traded companies are bought, sold, and issued. Such financial activities are carried out through institutionalised formal exchanges (whether physical or electronic) or through over-the-counter marketplaces that follow a set of rules. The stock market brings together, interacts with, and transacts with many buyers and sellers of securities. Stock markets enable the price discovery of corporate shares and serve as a barometer for the overall economy. Because of the large number of stock market participants, one can often be assured of a fair price and a high degree of liquidity because various market participants compete for the best price. Stock markets provide a safe and regulated environment where market participants can confidently trade shares and other eligible financial instruments with zero to low operational risk.

Studying historical market data, such as price and volume, is technical analysis. Technical analysts aim to predict future market behaviour using market psychology, behavioural economics, and quantitative analysis. Chart patterns and technical (statistical) indicators are the two most common types of technical analysis. Technical analysis is a catch-all term for several strategies that interpret price action in a stock. The majority of technical analysis is concerned with determining whether a current trend will continue and, if not when it will reverse. Some technical analysts swear by trendlines, while others swear by candlestick formations, and still, others swear by bands and boxes created through mathematical visualisation. To identify potential entry and exit points for trades, most technical analysts employ various tools. The core principle underlying technical analysis is that the market price reflects all available information that could impact a market. As a result, there's no need to look at economic, fundamental, or new developments since they're already priced into a given security. Technical analysts generally believe that prices move in trends and history tends to repeat itself when it comes to the market's overall psychology. The two major types of technical analysis are chart patterns and technical (statistical) indicators. Chart patterns are a subjective form of technical analysis where technicians attempt to identify areas of support and resistance on a chart by looking at specific patterns. These patterns, underpinned by psychological factors, are designed to predict where prices are headed, following a breakout or breakdown from a specific price point and time. Technical indicators are a statistical form of technical analysis where technicians apply various mathematical formulas to prices and volumes. The most common technical indicators are moving averages, which smooth price data to help make it easier to spot trends.

In this article, we'll explore various datasets of share prices of 6 firms and also see how the value of INR has fluctuated over time with respect to USD. We'll utilise mathematical models like regression, moving average and the ARIMA model to estimate important quantities that shall help a trader in the market make better decisions. This paper focuses on the mathematical aspect used to analyse the stock market data in Sec. II. In Sec. III, the concentration is on the given data and performing in-depth analysis to understand the behaviour of the stock market. Finally, Sec. IV concludes all the insights and findings from the study and the areas for further research.

## II. Mathematics behind Time Series Analysis

### A. Dataset Terminologies

- *Open*: The price at which the first share is traded on a specific trading day.
- *High*: The highest price at which a share is traded on a specific trading day.
- *Low*: The lowest price at which a share is traded on a specific trading day.
- *Close*: The price at which the last share is traded on a specific trading day.
- *Volume*: The number of shares exchanged in a specific trading day.

### B. Moving Averages

A moving average is a statistical calculation used to analyse data points by calculating a series of averages of different subsets of the entire data set. A moving average (MA) is a stock indicator commonly used in technical analysis in finance. The purpose of calculating a stock's moving average is to smooth out price data by creating a constantly updated average price.

The effects of random, short-term fluctuations on the price of a stock over a specified time frame are mitigated by calculating the moving average. Moving average is a simple, technical analysis tool. Moving averages are usually calculated to identify the trend direction of a stock or to determine its support and resistance levels. It is a trend-following—or lagging—indicator because it is based on past prices. The longer the time period for the moving average, the greater the lag. So, a 200-day moving average will have a much greater degree of lag than a 20-day MA because it contains prices for the past 200 days. Moving averages are a totally customizable indicator, which means that an investor can freely choose whatever time frame they want when calculating an average. The most common time periods used in moving averages are 10, 20, 25, 50, 100, and 200 days. The shorter the time span used to create the average, the more sensitive it will be to price changes. The longer the time span, the less sensitive the average will be. Investors may choose different time periods of varying lengths to calculate moving averages based on their trading objectives. Shorter moving averages are typically used for short-term trading, while longer-term moving averages are more suited for long-term investors.

A rising moving average indicates that the security is in an uptrend, while a declining moving average indicates that it is in a downtrend. Similarly, upward momentum is confirmed with a bullish crossover, which occurs when a short-term moving average crosses above a longer-term moving average. Conversely, downward momentum is confirmed with a bearish crossover, which occurs when a short-term moving average crosses below a longer-term moving average.

**Simple Moving Average**: The simplest form of a moving average, known as a simple moving average (SMA), is calculated by taking the arithmetic mean of a given set of values over a specified period of time. In other words, a set of numbers–or prices in the case of financial instruments–are added together and then divided by the number of prices in the set. The formula for calculating the simple moving average of a security is as follows:

$$SMA = \frac{A_1 + A_2 + \ldots + A_n}{n} \tag{1}$$

where
A = Average in period n
n = Number of time periods

### C. ARIMA Model

A famous and widely used forecasting method for time-series prediction is the AutoRegressive Integrated Moving Average (ARIMA) model. ARIMA models are capable of capturing a suite of different standard temporal structures in time-series data. We can break down the abbreviation as follows:

- *Auto Regressive (AR)*: the model uses the dependent relationship between an observation and some predefined number of lagged observations (also known as "time lag" or "lag").
- *Integrated (I)*: the model employs differencing of raw observations (e.g. it subtracts an observation from an observation at the previous time step) in order to make the time-series stationary.MA:
- *Moving Average (MA)*: the model exploits the relationship between the residual error and the observations.

Considering the attributes mentioned above, the ARIMA(p,d,q) model has the following 3 hyperparameters:

- p: number of lag observations.
- d: the degree of differencing.
- q: the size/width of the moving average window.

To implement the ARIMA model, we'll make use of the *statsmodels* library in python. To visualise and benchmark its performance, we'll utilise the *matplotlib* library and *root mean squared error* metric.

## III. Time Series Analysis on real-life datasets

This section will have 7 subsections, each one for each of the 7 datasets, which include stock prices of 3 IT companies, stock prices of 3 banking institutions and USD-INR exchange rate over the last 3 years. Briefly, I will first performing Exploratory Data Analysis to derive insights and features from the trends observed in the date. Subsequently, I will build predictive models to predict one or more quantities and rate the quality of the models with descriptive metrics. Finally, we'll try to see if there's any correlation between the stocks of different companies within the same industry.

### A. Cognizant (2 Jan 2019 - 1 Oct 2021)

Cognizant is an American multinational information technology services and consulting company. It is headquartered in Teaneck, New Jersey, United States. The data we have spans nearly 3 years of stock prices. We learn some interesting insights from the EDA we perform, which are summarised in the next section.

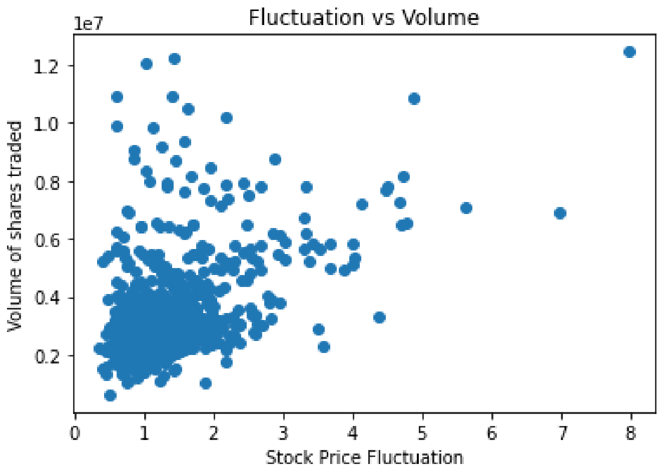


Fig. 1. Correlation between Fluctuation and Volume

*1) Exploratory Data Analysis:* There aren't any null values in the dataset. In total there are 694 rows and 6 columns. Main insights from this data:

- Extreme volume outlier: On 3 Oct 2019, a disproportionately large volume of trades was observed (of the order $4x10^7$). It is observed that a severe 20% price drop was also observed simultaneously, implying that price fluctuations may have a correlation with volume traded.
- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed, implying that there are more trades in the market when price of the stock varies more. This can be seen in Fig. 1.
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 2
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 3, we can see that the 50-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly low likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. 4 and Fig. 5 summarise the distribution of returns

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 6 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The hyperparameters chosen for the best fit are:

1) n_estimators: 400

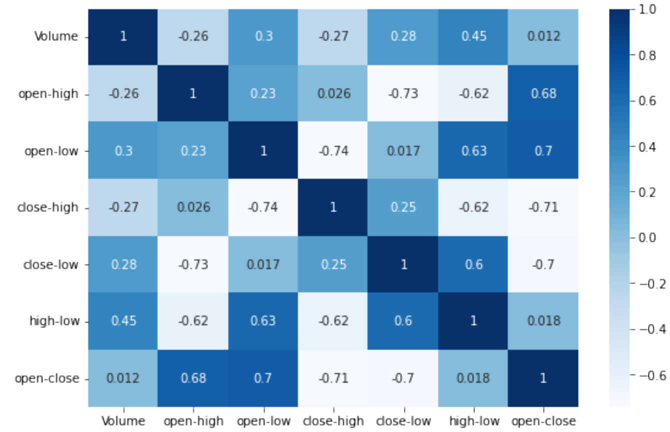


Fig. 2. Correlation Heatmap for Cognizant Data

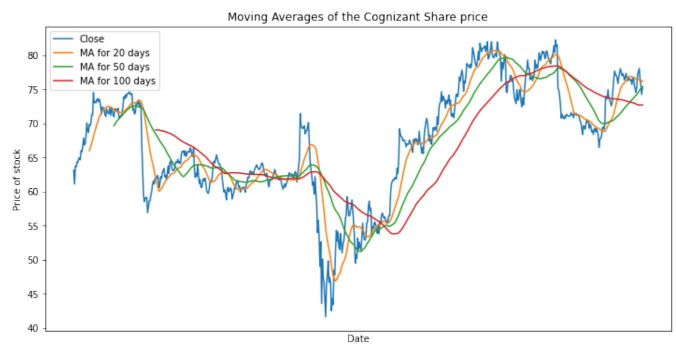


Fig. 3. Moving Average Plots for Cognizant

2) min_samples_split: 10
3) min_samples_leaf: 4
4) max_features: 'sqrt'
5) max_depth: 90
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Fig. 7 is an autocorrelation plot with a lag of 3. We see a strong positive correlation and can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 8.

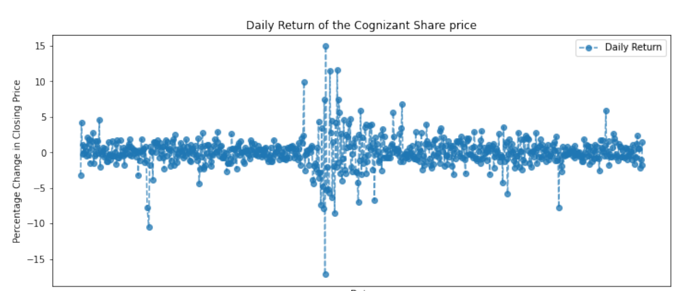


Fig. 4. Plot Daily Returns over time

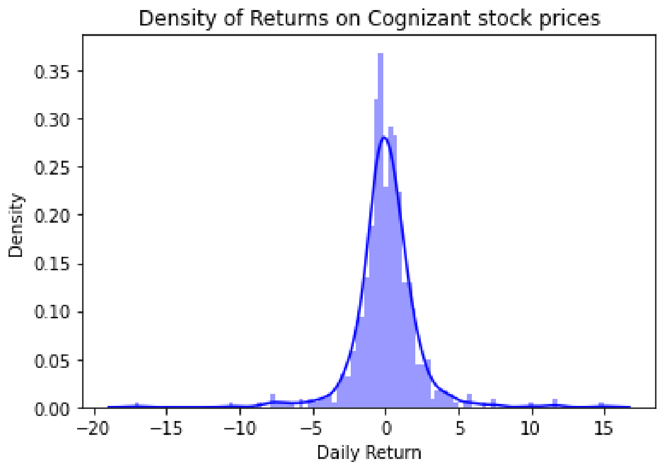


Fig. 5. Density Distribution of Daily Returns

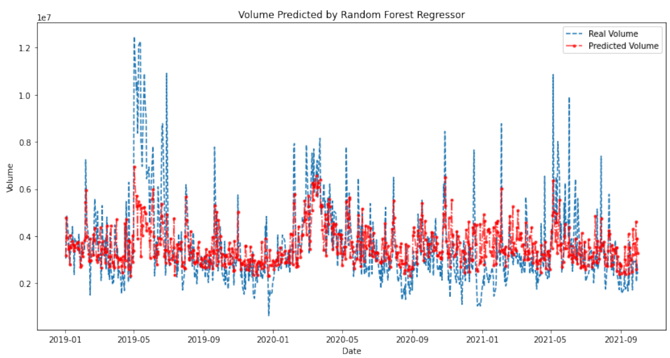


Fig. 6. Volume Prediction of Cognizant Stocks

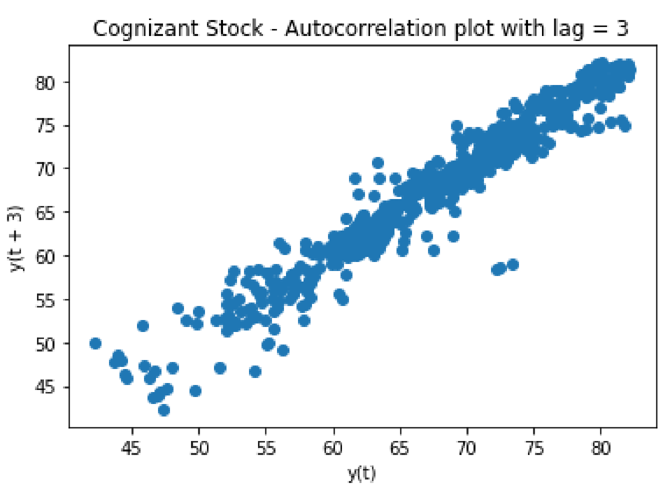


Fig. 7. Autocorrelation estimation

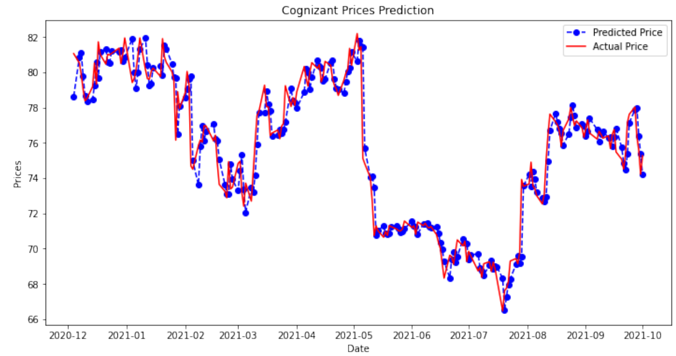


Fig. 8. ARIMA's Cognizant Share price prediction on Test Data

| Estimated Quantity | Model | Train Error | Test Error |
|---|---|---|---|
| Volume | RF Regressor | 1.27e+03 | 1.36e+03 |
| Price | ARIMA | - | 1.27 |

TABLE I
PERFORMANCE REPORT ON PREDICTIVE MODELS

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that our model fits reasonable well. As for Stock Price precition, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

### *B. HCL Technologies (5 Oct 2020 - 4 Oct 2021)*

HCL Tech is an Indian multinational information technology services and consulting company headquartered in Noida, Uttar Pradesh. We learn some interesting insights from the EDA we perform, which are summarised in the next section.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 248 non-null rows and 6 columns. Main insights from this data:

- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed, implying that there are more trades in the market when price of the stock varies more
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 9
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 10, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly nominal likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. 11 and Fig. 12 summarise the distribution of returns

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The

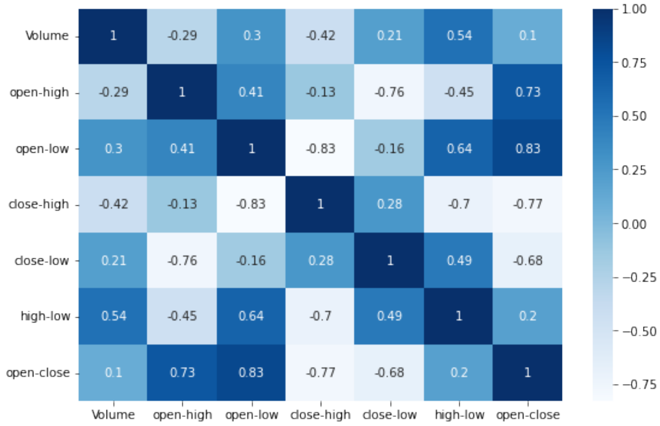


Fig. 9. Correlation Heatmap for HCL Tech Data

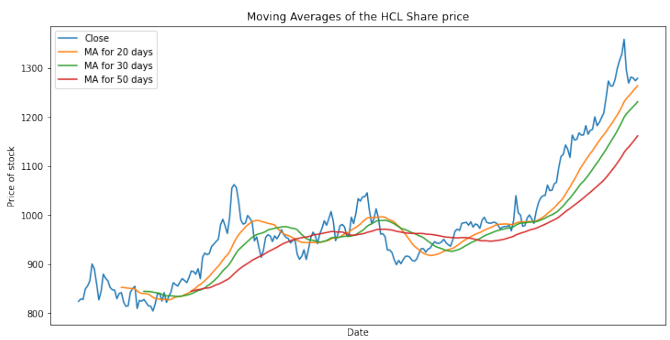


Fig. 10. Moving Average Plots for HCL Tech

RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 13 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The hyperparameters chosen for the best fit are:

1) n_estimators: 800
2) min_samples_split: 10
3) min_samples_leaf: 4
4) max_features: 'sqrt'

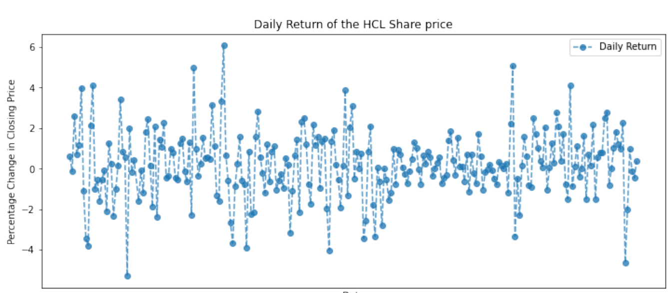


Fig. 11. Plot of Daily Returns over time

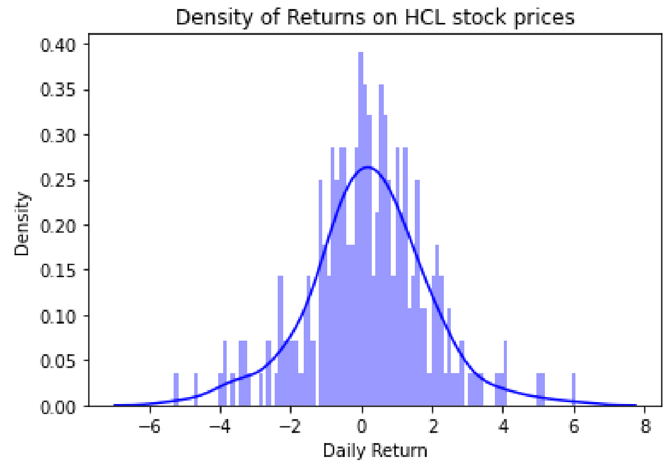


Fig. 12. Density Distribution of Daily Returns

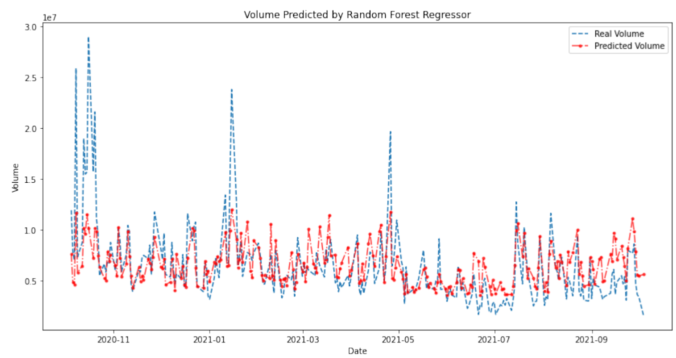


Fig. 13. Volume Prediction of HCL Tech Stocks

5) max_depth: 100
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 14.

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that our model fits reasonable well. As for Stock Price precition, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

| **Estimated Quantity** | **Model** | **Train Error** | **Test Error** |
|---|---|---|---|
| Volume | RF Regressor | 2.43e+03 | 3.61e+03 |
| Price | ARIMA | - | 335.99 |

TABLE II
PERFORMANCE REPORT ON PREDICTIVE MODELS

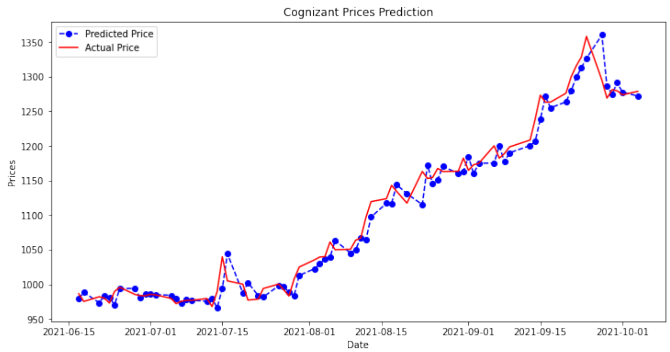


Fig. 14. ARIMA's HCL Tech Share price prediction on Test Data

### *C. Infosys (5 Oct 2020 - 4 Oct 2021)*

Infosys Limited is an Indian multinational information technology company that provides business consulting, information technology and outsourcing services. The company was founded in Pune and is headquartered in Bangalore.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 248 non-null rows and 6 columns. Main insights from this data:

- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed, implying that there are more trades in the market when price of the stock varies more
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 15
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 16, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly high likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. 17 and Fig. 18 summarise the distribution of returns

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 19 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The hyperparameters chosen for the best fit are:

1) n_estimators: 600
2) min_samples_split: 2
3) min_samples_leaf: 4

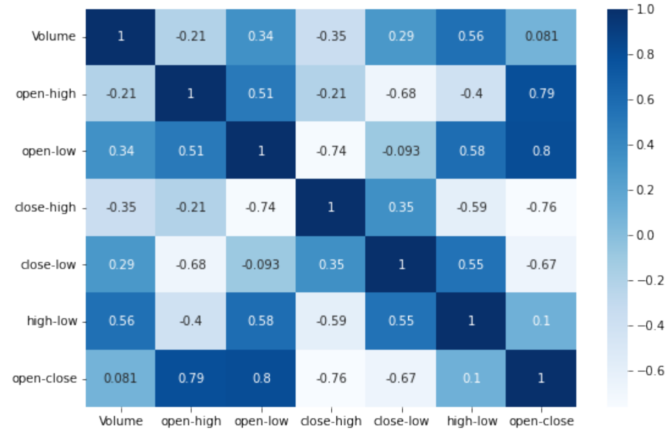


Fig. 15. Correlation Heatmap for Infosys Data

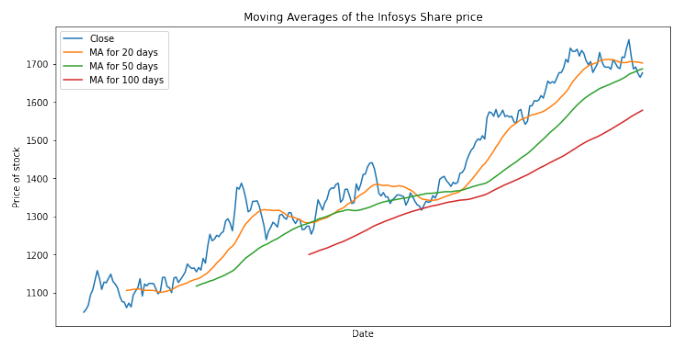


Fig. 16. Moving Average Plots for Infosys

4) max_features: 'auto'
5) max_depth: 40
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 20.

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that

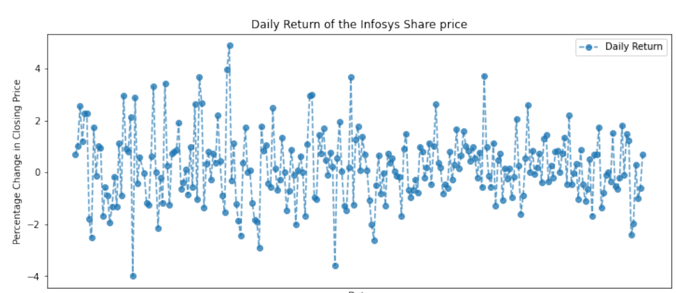


Fig. 17. Plot of Daily Returns over time

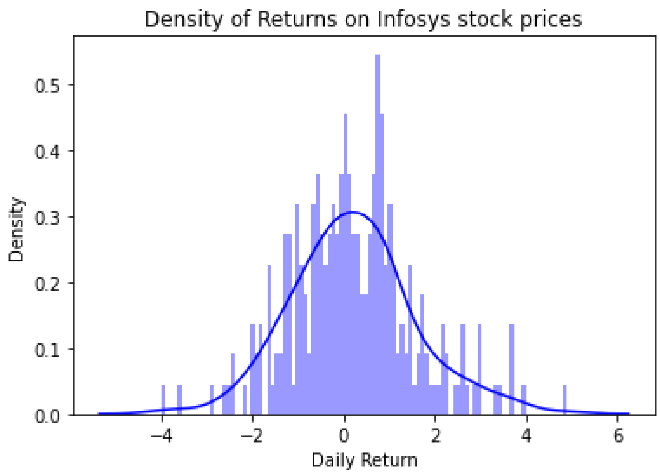


Fig. 18. Density Distribution of Daily Returns

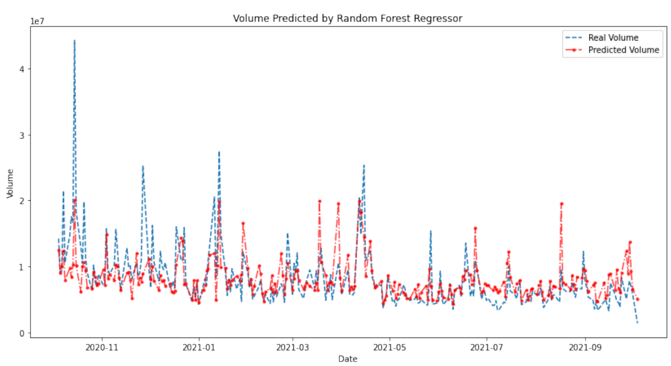


Fig. 19. Volume Prediction of Infosys Stocks

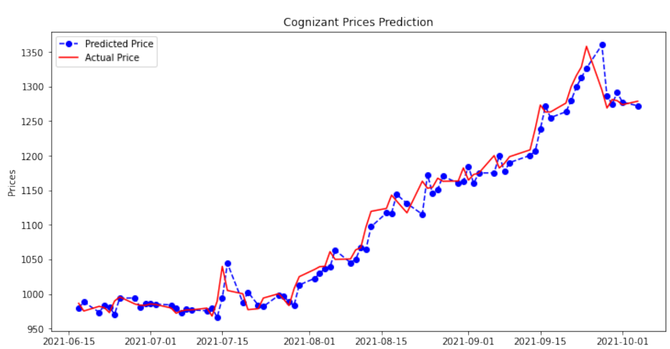


Fig. 20. ARIMA's Infosys Share price prediction on Test Data

| Estimated Quantity | Model | Train Error | Test Error |
|---|---|---|---|
| Volume | RF Regressor | 3.09e+03 | 4.36e+03 |
| Price | ARIMA | - | 319.88 |

TABLE III
PERFORMANCE REPORT ON PREDICTIVE MODELS

our model fits reasonable well. As for Stock Price precition, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

### D. All 3 tech companies

When we plot the pairplots of the prices of each tech stock as in Fig. 21, we notice that there is a strong positive correlation between the HCL stock's price and Infosys' stock price. This makes sense as both companies are based in India, whereas Cognizant is an MNC.

### E. HDFC (1 Jan 2019 - 1 Oct 2021)

HDFC Bank Limited is an Indian banking and financial services company headquartered in Mumbai. It is India's largest private sector bank by assets and world's 10th largest bank by market capitalisation as of April 2021.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 678 non-null rows and 6 columns. Main insights from this data:

- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed, implying that there are more trades in the market when price of the stock varies more
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 22
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 23, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly low likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. 24 and Fig. 25 summarise the distribution of returns.

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 19 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The hyperparameters chosen for the best fit are:

1) n_estimators: 400

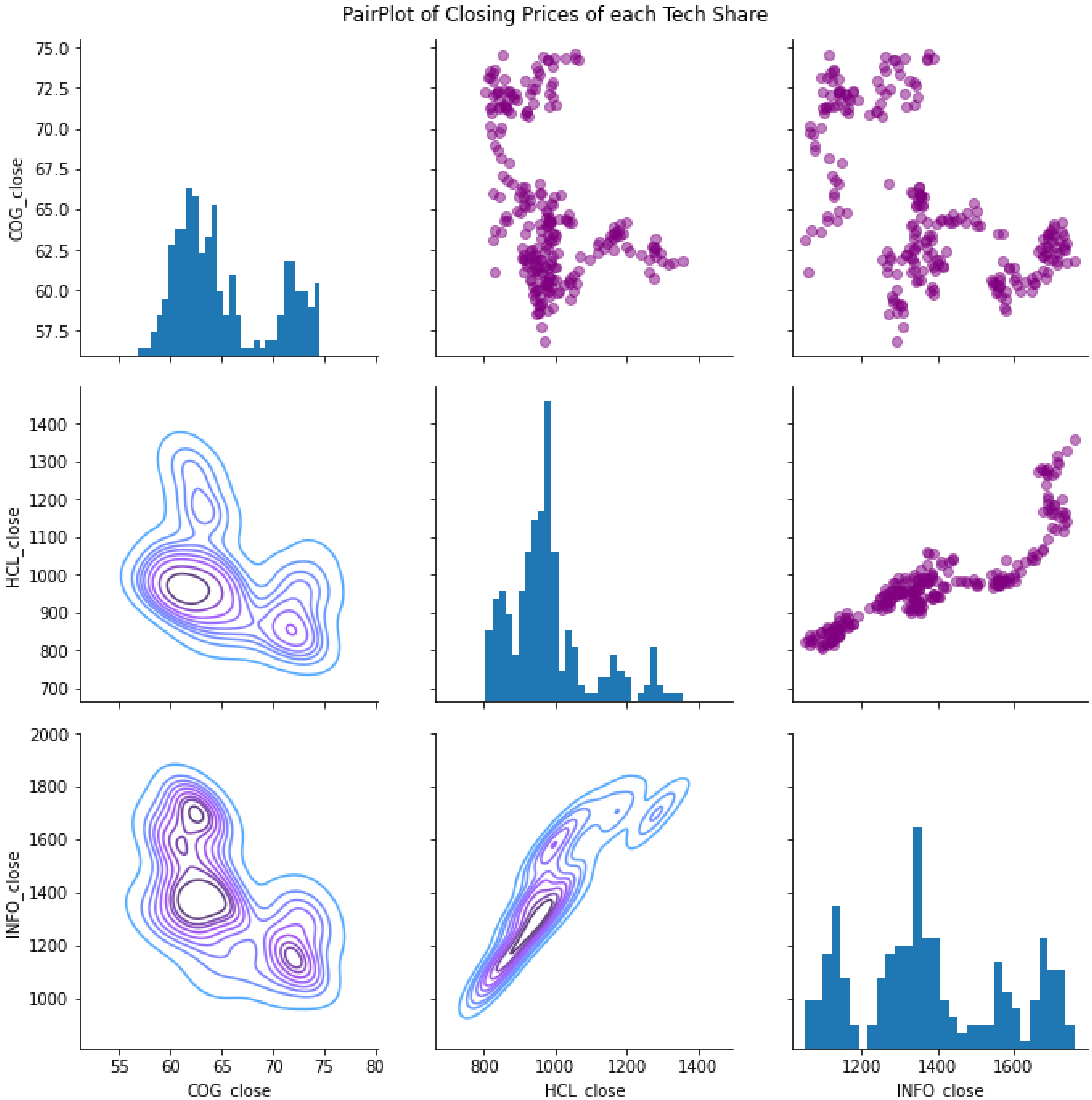


Fig. 21. Pairplots of Prices of the 3 tech commpanies

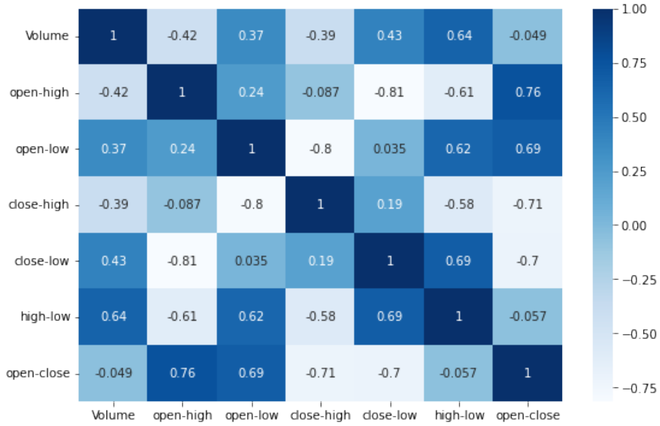


Fig. 22. Correlation Heatmap for HDFC Data

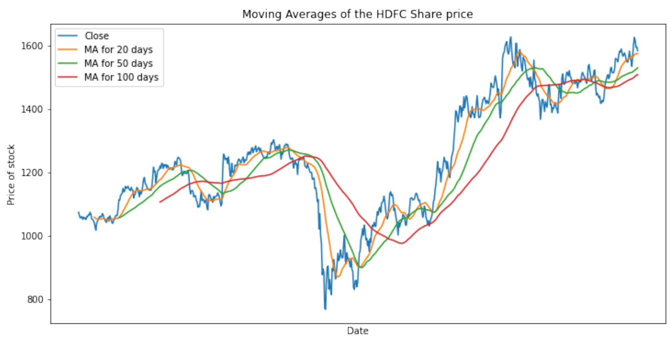


Fig. 23. Moving Average Plots for HDFC

2) min_samples_split: 5
3) min_samples_leaf: 4
4) max_features: 'sqrt'
5) max_depth: 70
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 27.

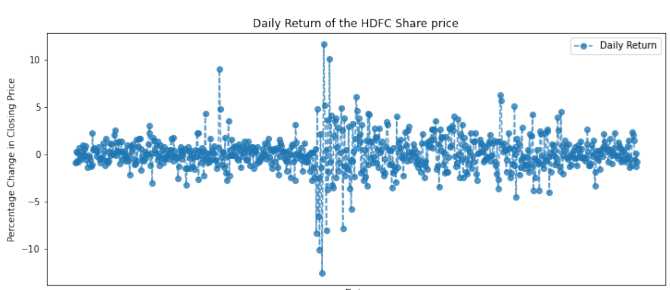


Fig. 24. Plot of Daily Returns over time

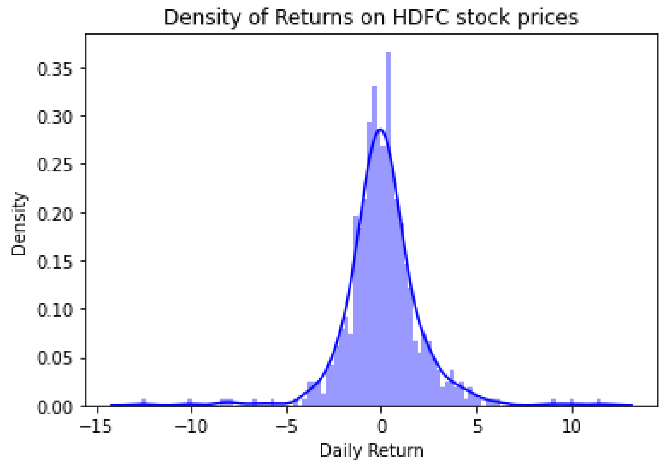


Fig. 25. Density Distribution of Daily Returns

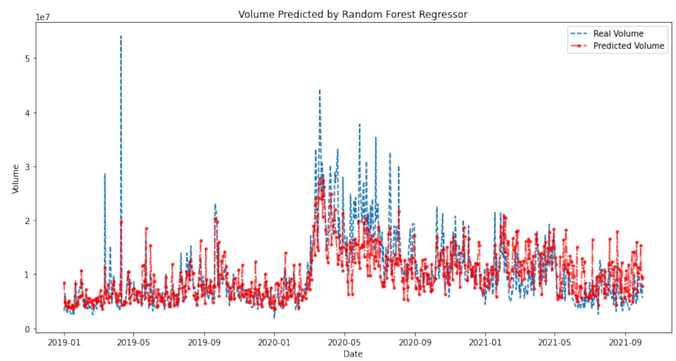


Fig. 26. Volume Prediction of HDFC Stocks

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that our model fits reasonable well. As for Stock Price prediction, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

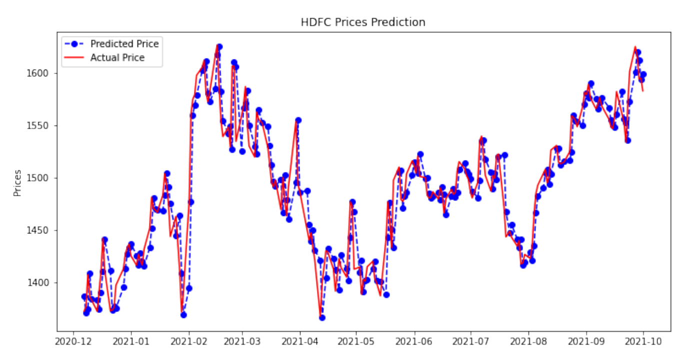


Fig. 27. ARIMA's HDFC Share price prediction on Test Data

| Estimated Quantity | Model | Train Error | Test Error |
|---|---|---|---|
| Volume | RF Regressor | 3.58e+03 | 5.4e+03 |
| Price | ARIMA | - | 537.88 |

TABLE IV
PERFORMANCE REPORT ON PREDICTIVE MODELS

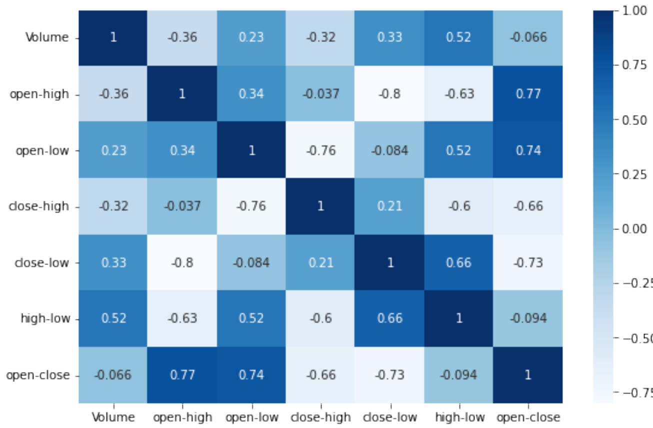


Fig. 28. Correlation Heatmap for ICICI Data

### F. ICICI (1 Jan 2019 - 1 Oct 2021)

ICICI Bank Limited is an Indian Private bank. It is headquartered at Mumbai.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 678 non-null rows and 6 columns. Main insights from this data:

- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed, implying that there are more trades in the market when price of the stock varies more
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 28
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 29, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly low likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. **??** and Fig. 31 summarise the distribution of returns.

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 19 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The hyperparameters chosen for the best fit are:

1) n_estimators: 1000
2) min_samples_split: 10
3) min_samples_leaf: 2
4) max_features: 'sqrt'
5) max_depth: 10
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag

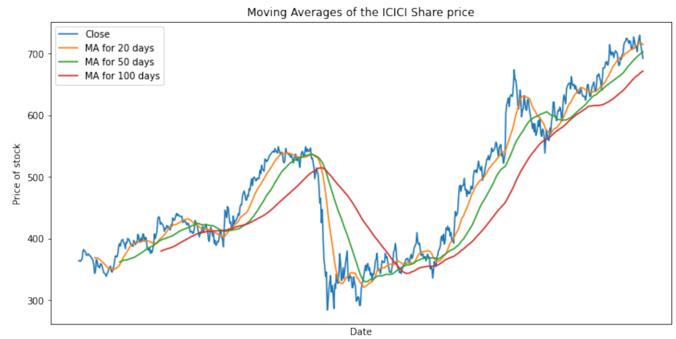


Fig. 29. Moving Average Plots for ICICI

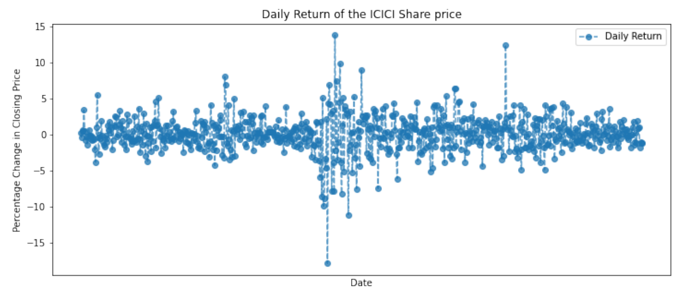


Fig. 30. Plot of Daily Returns over time

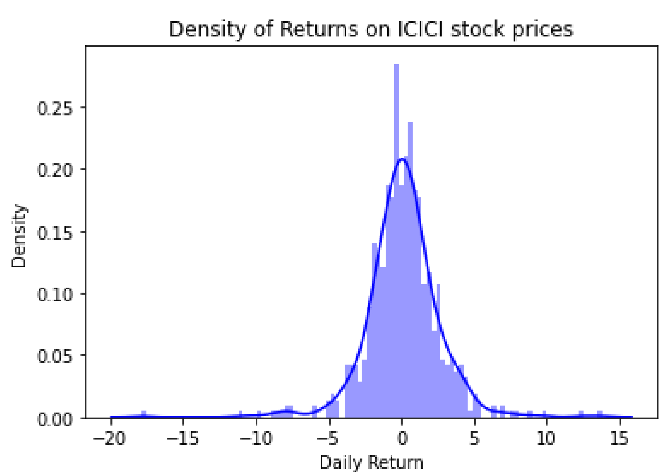


Fig. 31. Density Distribution of Daily Returns

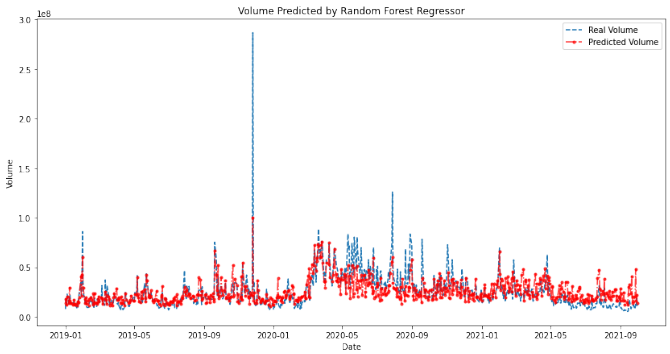


Fig. 32. Volume Prediction of ICICI Stocks

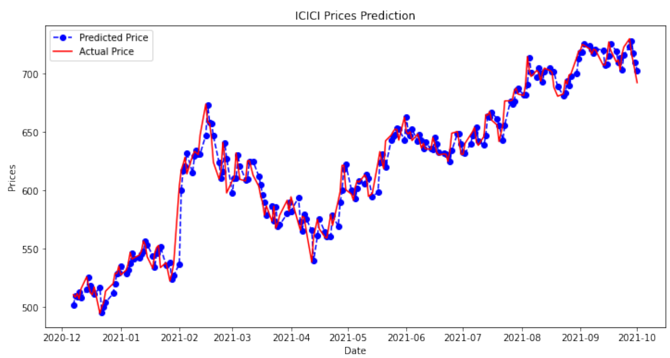


Fig. 33. ARIMA's ICICI Share price prediction on Test Data

of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 33.

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that our model fits reasonable well. As for Stock Price prediction, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

### *G. SBI (1 Jan 2019 - 1 Oct 2021)*

State Bank of India is an Indian multinational public sector bank and financial services statutory body headquartered in Mumbai, Maharashtra.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 678 non-null rows and 6 columns. Main insights from this data:

- Fluctuation vs Volume: On plotting fluctuations (high-low) vs volume, a loosely positive correlation is observed,

| Estimated Quantity | Model | Train Error | Test Error |
|---|---|---|---|
| Volume | RF Regressor | 1.22e+04 | 1.30e+04 |
| Price | ARIMA | - | 131.36 |

TABLE V
PERFORMANCE REPORT ON PREDICTIVE MODELS

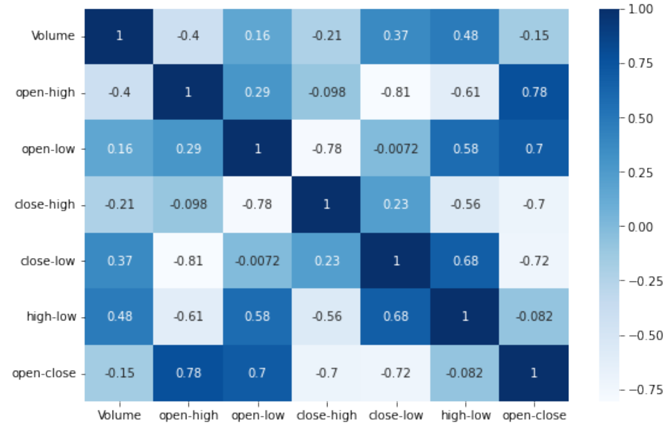


Fig. 34. Correlation Heatmap for SBI Data

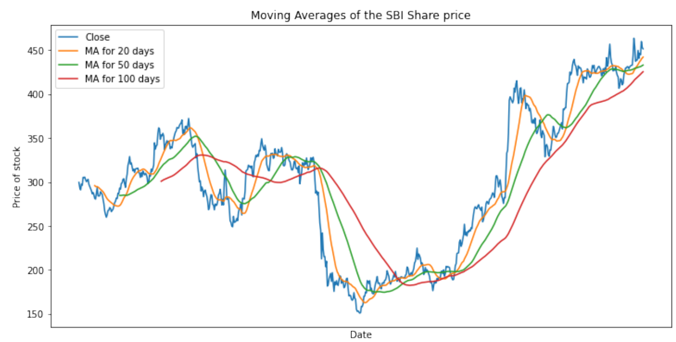


Fig. 35. Moving Average Plots for SBI

implying that there are more trades in the market when price of the stock varies more
- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 34
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 35, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly low likelihood of running into high returns with this share. This can be quantified by estimating the kurtosis of the density of daily returns. Fig. 43 and Fig. 37 summarise the distribution of returns.

*2) Predictive Modelling of Volume and Share price:* Considering Volume is a continuous variable, I employed a Random Forest Regressor to build a predictive model. The RandomForestRegressor class from the sklearn library makes fitting a model fairly simple and straightforward. I also tuned the hyperparameters using RandomSearchCV which randomly searches the hyperparameter space and evaluates models with k-fold cross validation. Fig. 19 represents how well our model estimates Volume. It seems to fit fairly well but is unable to capture sudden changes in volume as accurately. The

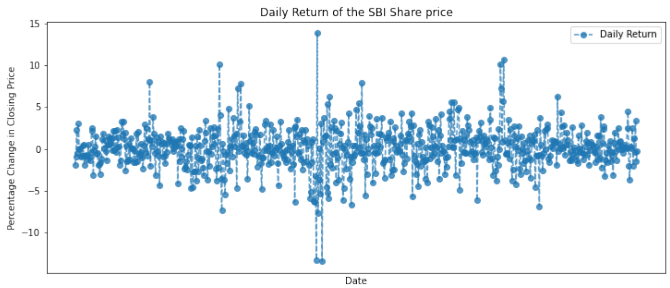


Fig. 36. Plot of Daily Returns over time

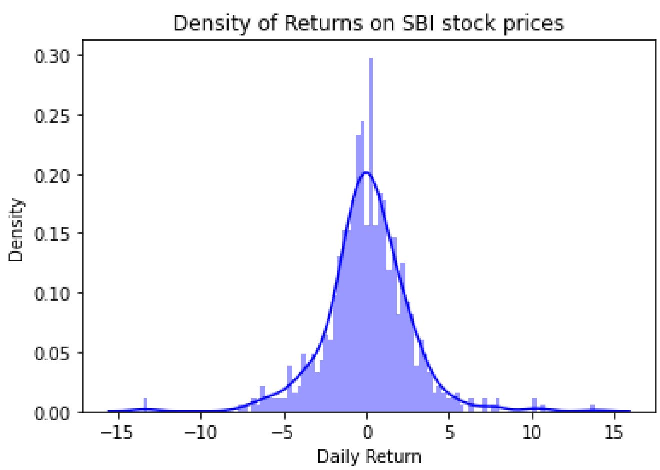


Fig. 37. Density Distribution of Daily Returns

hyperparameters chosen for the best fit are:

1) n_estimators: 400
2) min_samples_split: 2
3) min_samples_leaf: 4
4) max_features: 'sqrt'
5) max_depth: 10
6) bootstrap: True

To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag

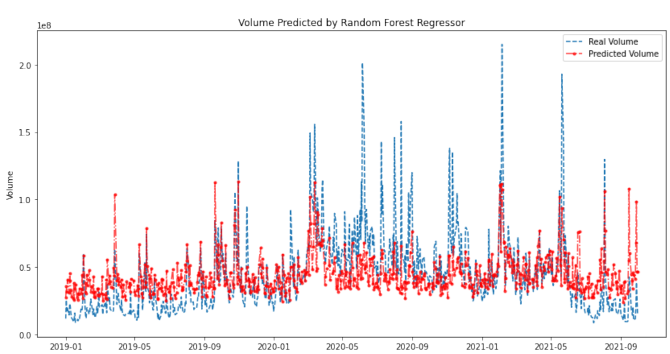


Fig. 38. Volume Prediction of SBI Stocks

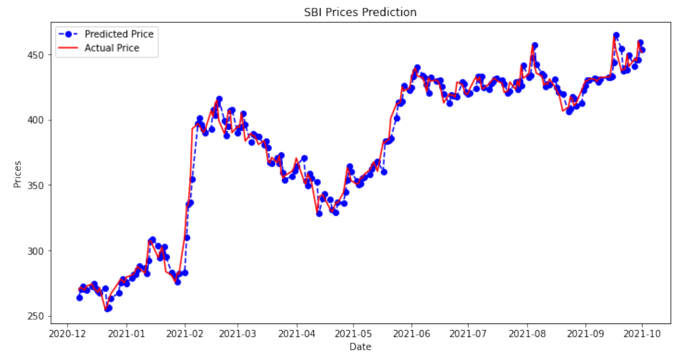


Fig. 39. ARIMA's SBI Share price prediction on Test Data

| Estimated Quantity | Model | Train Error | Test Error |
|---|---|---|---|
| Volume | RF Regressor | 1.85e+04 | 2.76e+04 |
| Price | ARIMA | - | 537.88 |

TABLE VI
PERFORMANCE REPORT ON PREDICTIVE MODELS

of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 39.

*3) Performance of Predictive Models:* Although the error on Volume prediction seems high, the earlier plot shows that our model fits reasonable well. As for Stock Price prediction, the ARIMA model fits exceptionally well with a very low test set error. Train error has not been reported for ARIMA as it will anyways be low.

### *H. All 3 bank companies*

When we plot the pairplots of the prices of each tech stock as in Fig. 40, we notice that there is a strong positive correlation between all 3 stock prices. This makes sense as both companies are based in India.

### *I. USD-INR Exchange Rate (1 Jan 2019 - 4 Oct 2021)*

USD is United States Dollar and INR is Indian National Ruppee. We'll be analysing the exchange rate in this section.

*1) Exploratory Data Analysis:* There aren't many null values in the dataset. Rows that are null are dropped. In total there are 699 non-null rows and 6 columns. Main insights from this data:

- On generating 6 additional features, namely open-high, open-low, close-high, close-low, high-low and open-close, some reasonable high correlations are observed as seen in Fig. 41
- On plotting Moving Averages of the Cognizant Share Price as in Fig. 42, we can see that the 20-days moving average best describes the trend, while simultaneously catching the price drops.
- On performing Daily Returns analysis, we can see that there is a fairly nominal likelihood of running into high returns with this share. This can be quantified by

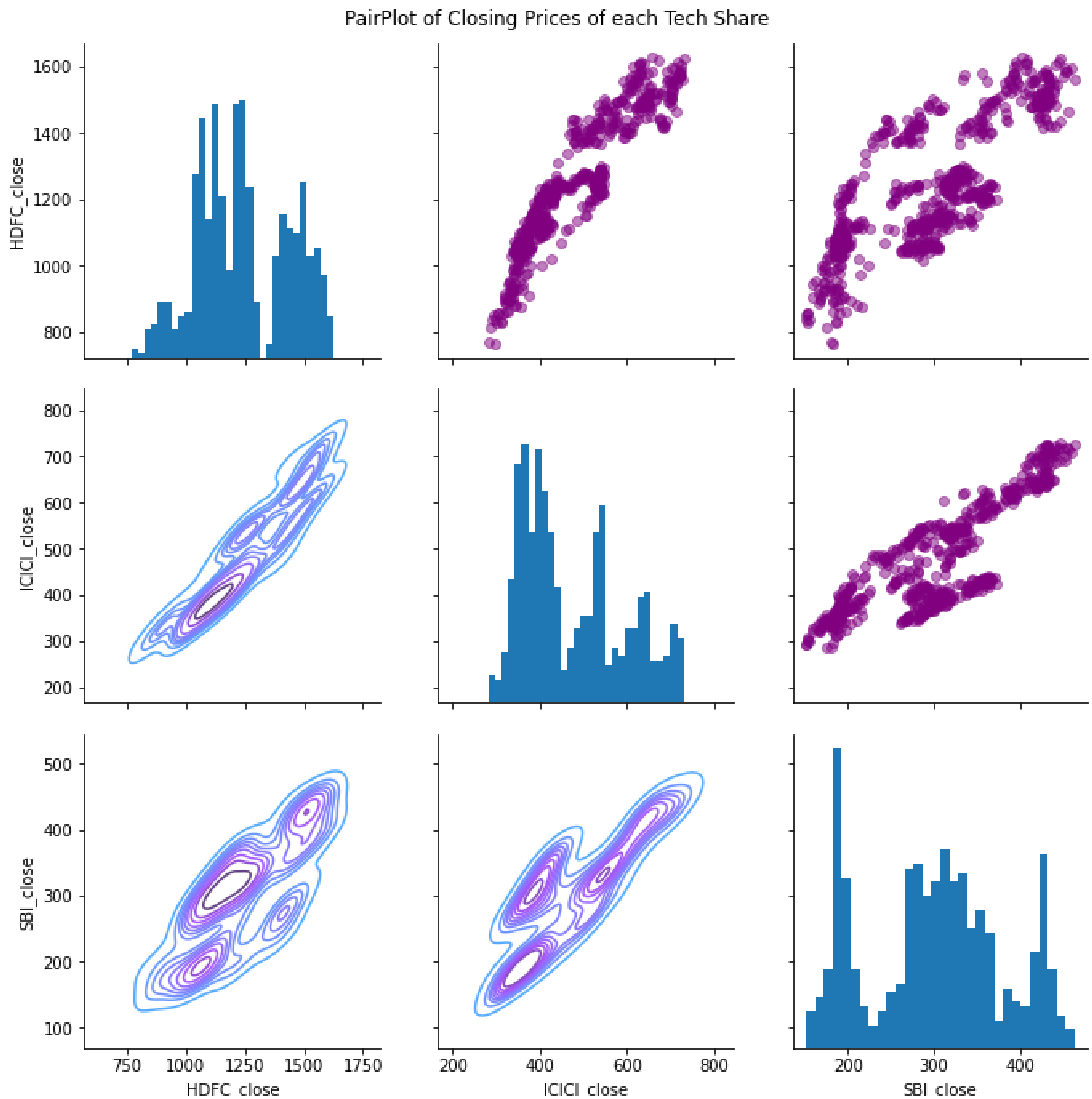


Fig. 40. Pairplots of Prices of the 3 bank commpanies

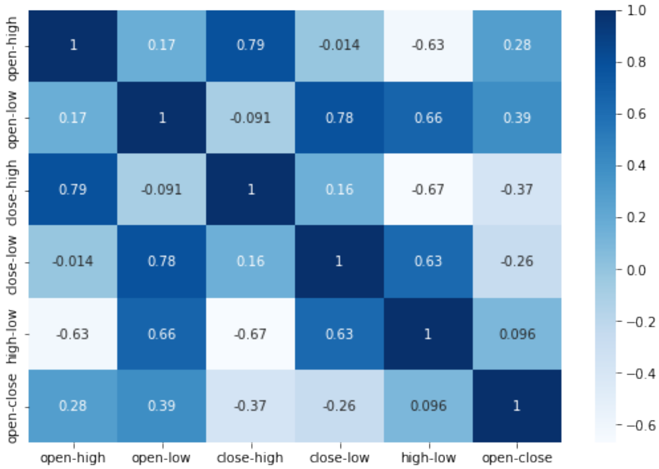


Fig. 41. Correlation Heatmap for USD-INR Exchange Rate Data

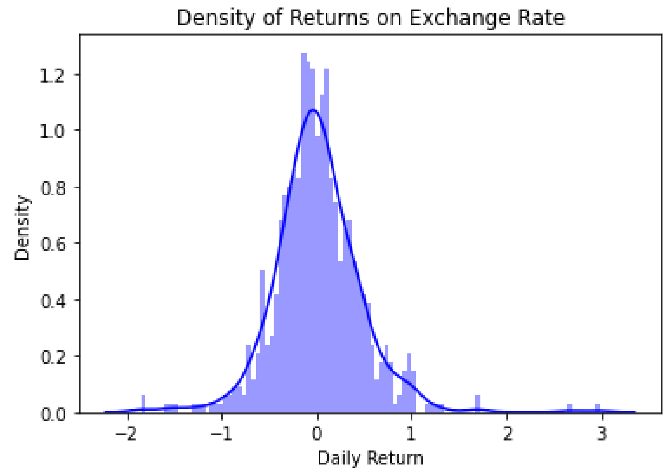


Fig. 44. Density Distribution of Daily Returns

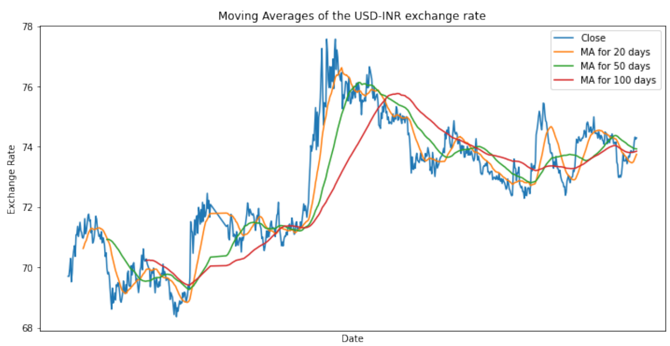


Fig. 42. Moving Average Plots for USD-INR Exchange Rate

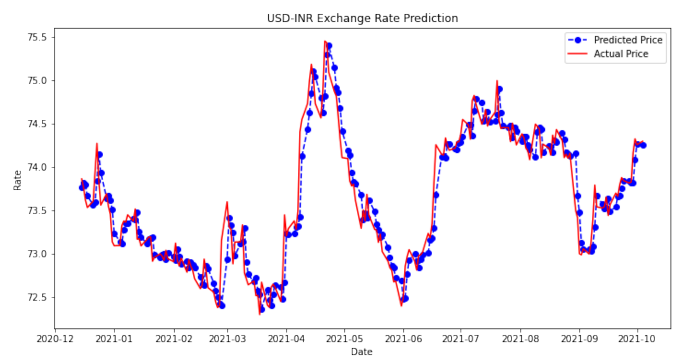


Fig. 45. ARIMA's USD-INR Exchange Rate prediction on Test Data

estimating the kurtosis of the density of daily returns. Fig. **??** and Fig. 44 summarise the distribution of returns.

*2) Predictive Modelling of USD-INR Exchange Rate:* To build a time series predictive model to estimate share price, I built an ARIMA model. Firtly, we have to check if the data is autocorrelated. Plotting autocorrelation with a lag of 3 shows a strong positive correlation, implying that can hence go ahead with fitting an ARIMA model with the (p, d, q) parameters of (4,1,0). The ARIMA model fits the test data very well, as can be seen in Fig. 45.

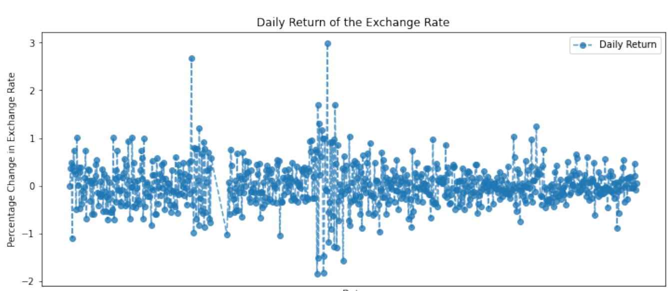


Fig. 43. Plot of Daily Returns over time

*3) Performance of Predictive Models:* With a prediction error of 0.053, the ARIMA model fits exceptionally well with a very low test set error.

## IV. Conclusion

We've analysed in detail how each stock fluctuated over time. We also saw the impact of the USD-INR exchange rate on each stock. Random Forest Regressor proved to be an efficient algorithm to predict continuous outputs. ARIMA model performed exceptionally well on time series data with autocorrelation.

## References

[1] https://towardsdatascience.com/time-series-forecasting-predicting-stock-prices-using-an-arima-model-2e3b3080bd70#:~text=A%20famous%20and%20widely%20used,structures%20in%20time%2Dseries%20data.

[2] Deisenroth, Marc Peter, Faisal, A. Aldo and Ong, Cheng Soon. Mathematics for Machine Learning. : Cambridge University Press, 2020, pp.370–380

[3] Boyd, Stephen, and Vandenberghe, Lieven, Introduction to Applied Linear Algebra, Cambridge University Press, 2018

[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013, pp.344–355

[5] https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/