Final Exam: An essay on Stock Analysis

1st Mihir Singh
Chemical Engineering
Indian Institute of Technology Madras
Chennai, India
ch20b065@smail.iitm.ac.in

***Abstract*—A stock exchange serves as a marketplace where individuals can buy and sell shares of publicly traded companies. It offers a platform facilitating seamless stock transactions conducted online. This study employs various forecasting software and methodologies to predict the stock prices of several organizations, and computational techniques to interpret any relationships that might exist between these stock prices . The analysis covers the stock markets of Cognizant, HDFC, HCL, Infosys, SBI, and ICICI for the period spanning 2019 to 2021.**

I. Introduction

A stock represents a financial instrument that signifies a proportional stake in a company's assets and profits. Stocks are alternatively referred to as shares or equity in a corporation. Our current economic landscape is characterized by uncertainty, where a company experiencing significant profitability today might face challenges in the future. Given the substantial returns it offers, the stock market is presently capturing widespread attention across various sectors. In recent years, computerized trading has taken center stage, with algorithms playing a crucial role in making split-second trading decisions.

One of these models is the ARIMA model which stands for AutoRegressive Integrated Moving Average, is a popular and widely used time series forecasting model. The key components of ARIMA are:

- **AutoRegressive (AR) Term:** The AR term involves modeling the relationship between an observation and several lagged observations (previous time points).
- **Integrated (I) Term:** The I term represents the differencing of the raw observations to make the time series stationary. Stationarity is often a prerequisite for time series modeling.
- **Moving Average (MA) Term:** The MA term involves modeling the relationship between an observation and a residual error from a moving average model applied to lagged observations.

Deep learning has also been widely used in the stock market. Recurrent Neural Networks (RNNs) in the form of LSTMs are the most popular type of neural network utilized. LSTMs help solve the vanishing gradient problem in RNNs and are particularly effective in capturing long-range dependencies and patterns in sequential data, making them well-suited for tasks like time series forecasting.

Time series data regarding the stock prices of six different companies: Cognizant, HDFC, HCL, Infosys, SBI, and ICICI, and the US-INR exchange rate for the time periond spanning 2019 to 2021 is used for our analysis. A special focus is given to the stock price data of Cognizant to test out the performances of our forecasting techniques purely because it is the only data set with no missing values.

In addition to analysing the trends within each company, or more specifically cognizant, we also aim to see if there are any strong correlations between the performances of the various organization, and also whether these companies' performances in any way impact the US-INR exchange rate.

As we will be mostly dealing with regression tasks, the appropriate metrics need to also be used. These include the mean average error, root mean squared error, and the R-squared score.

II. ARIMA

AutoRegressive Integrated Moving Average (ARIMA) is a popular time series forecasting model that combines autoregression, differencing, and moving averages to capture and predict patterns in sequential data.

The key components of ARIMA are:

- **AutoRegressive (AR) Term:** The AR term involves modeling the relationship between an observation and several lagged observations (previous time points).
- **Integrated (I) Term:** The I term represents the differencing of the raw observations to make the time series stationary. Stationarity is often a prerequisite for time series modeling.
- **Moving Average (MA) Term:** The MA term involves modeling the relationship between an observation and a residual error from a moving average model applied to lagged observations.

The notation for ARIMA is typically written as ARIMA(p, d, q), where:

- **p**: The order of the AutoRegressive component.
- **d**: The degree of differencing.
- **q**: The order of the Moving Average component

The ARIMA model is specified based on the characteristics of the time series data. The process of determining the optimal values of p, d, and q involves analyzing the autocorrelation and partial autocorrelation functions of the time series.

The mathematical equation for ARIMA is as follows:

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)(1 - B)^d Y_t = c + (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)\varepsilon_t$$

Here:

- Y_t is the value of the time series at time t.
- B is the backshift operator, which represents the lag ($BY_t = Y_{t-1}$.
- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive parameters.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average parameters.
- d is the degree of differencing.
- c is a constant term.
- ε_t is the error term.

It's worth noting that this equation is a compact representation, and the actual values of ϕ_i and θ_i need to be estimated during the model training process using techniques like maximum likelihood estimation. The fitted ARIMA model can then be used for time series forecasting.

III. LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is specifically designed to address the vanishing gradient problem, which is a challenge associated with training traditional RNNs. LSTMs are particularly well-suited for modeling sequential data

The key features of an LSTM network are:

- **Memory Cell**: LSTMs use a memory cell that allows the network to store information over long sequences.
- **Gates**: There are three gates.
  - **Forget Gate**: Determines what information from the cell state to discard or keep.
  - **Input Gate**: Modifies the cell state with new information.
  - **Output Gate**: This prroduces the final output based on the modified cell state.
- **Hidden State**: LSTM networks maintain a hidden state that serves as a memory of the previous inputs and captures relevant information needed to make predictions. The hidden state is updated based on the current input and the information passed through the gates.

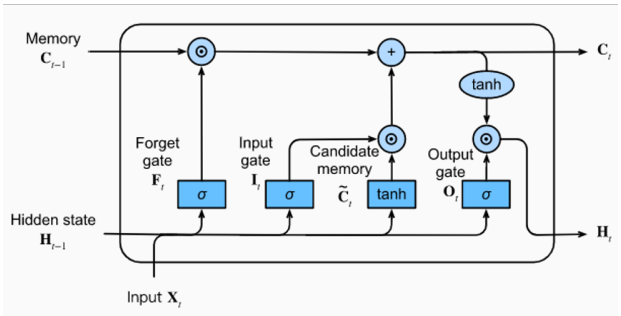

Fig. 1. LSTM unit

IV. THE PROBLEM

The data provided contains details about the stock prices of six different companies as well as the prices of the US-INR exchange rate.

A. Analyzing Cognizant

The company we way special focus to is Cognizant. Provided is the closing price, opening price, least price, highest price, and trade volume of the Cognizant stock from 2019 to 2021.

Fig. 2. Cognizant

First, to analyze the trend the closing prices of the stock is showing we look at the data's moving average plot taken over a period of 90 days or every financial quarter.

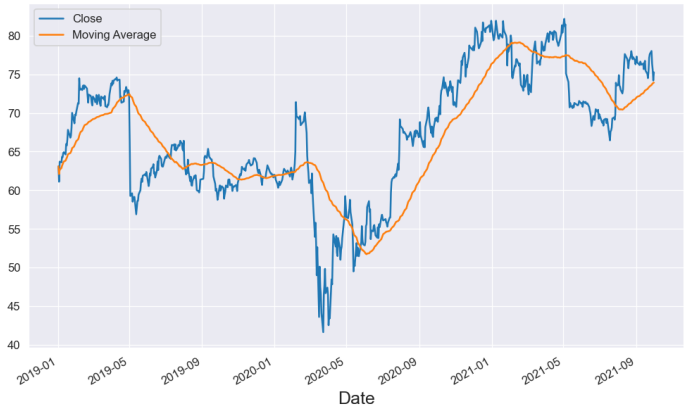

Fig. 3. Cognizant Closing Price Trend

Looking at Fig 3. we can see that the overall trend suggests that initially till the mid 2020s was the price of the stock was going down after which the stock suddenly increased in valuation before stagnating by the beginning of the next year 2021.

Next, we seek to understand any seasonal patterns that the time series may show. To do this, we observe the patterns the prices follow every month.

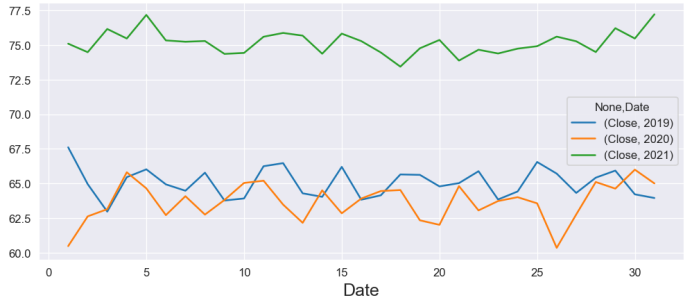

Fig. 4. Cognizant Seasonality

Fig 4. obviously seems to indicate that there are no concrete seasonal patterns in the data. This is further corroborated by the following seasonal deocomposition plot in Fig 5.

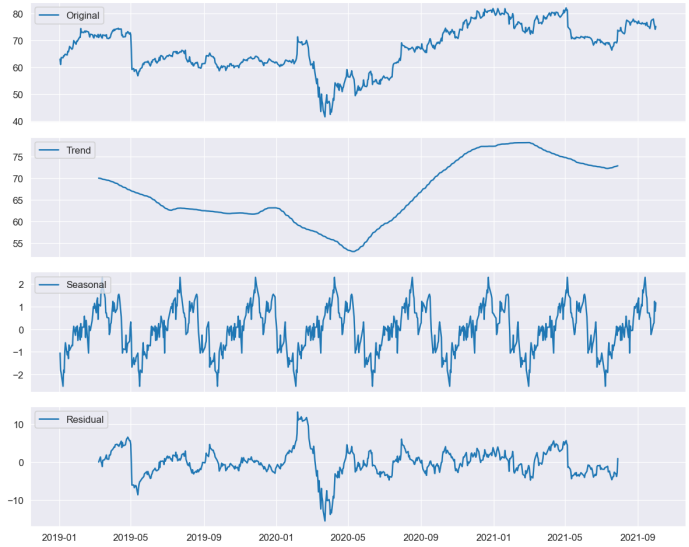

Fig. 5. Cognizant Seasonal Decomposition

Now that we have a rudimentary understanding of our data, we build an ARIMA model to assist us in forecasting the closing prices of the stock.

As mentioned earlier, the 3 parameters used by the ARIMA model are **p, d,** and **q**.

To determine **d**, we use the statistical test known as the **Augmented Dickey Fuller** Test. This test returned a **p-value** of **0.253171**. Since this is greater than the significance level of 0.05, we then difference the time series data and qualitatively analyse the corresponding auto-correlation plots.

These Auto-correlation plots can be seen in Fig 6.

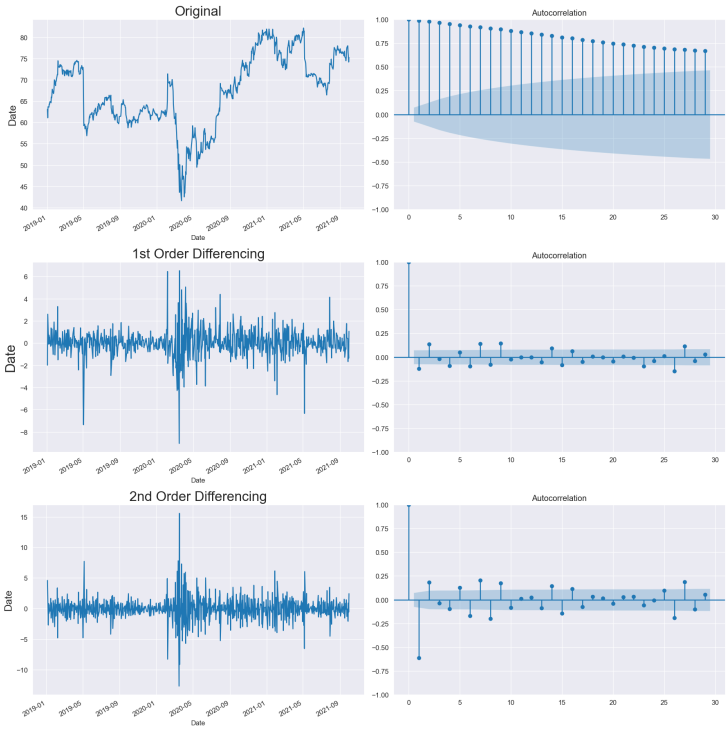

Fig. 6. auto-correlation plots for different levels of differencing

In the auto-correlation plot for 1st order differencing, we see the auto-correlation immediately reduce to within the significance level in the second step. This clearly indicates that the time series achieves stationarity when it is differenced by one order. This means that the most appropriate value for **d** is 1.

Next, we aim to identify the best value for the parameter **p**. To do so, we plot the **partial auto-correlation plot** of the one order differenced time series data.

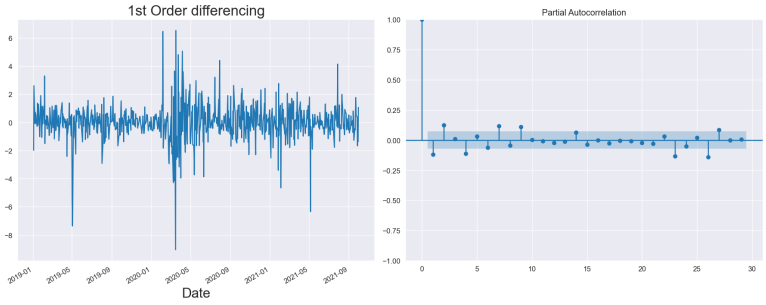

Fig. 7. Partial auto-correlation plot

Here, we see the partial auto-correlation plot fall to within the significance level of 0.05 by the first step. This implies that the most appropriate value for **p** is 1.

To find the best value of **q**, we repeat the process used find p but instead of using the partial auto-correlation plot, the regular auto-correlation plot is used.

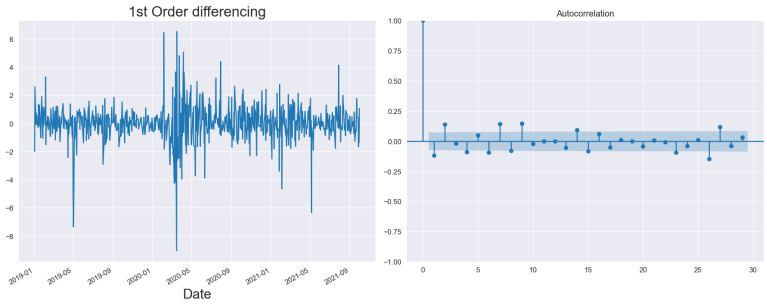

Fig. 8. Auto-correlation plot

Here also, we see the auto-correlation plot fall to within the significance level of 0.05 by the first step. This implies that the most appropriate value for **q** is 1.

Therefore, the ARIMA model used to analyse the closing prices of the Cognizant stock is **ARIMA(1,1,1)**.

The exact parameters and coefficients of the ARIMA model can be seen in the following figure (Fig 9.).

SARIMAX Results

Dep. Variable:	Close	No. Observations:	694
Model:	ARIMA(1, 1, 1)	Log Likelihood	-1154.275
Date:	Sat, 25 Nov 2023	AIC	2314.551
Time:	14:25:50	BIC	2328.174
Sample:	0	HQIC	2319.819
	- 694		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4826	0.148	-3.256	0.001	-0.773	-0.192
ma.L1	0.3521	0.155	2.275	0.023	0.049	0.656
sigma2	1.6377	0.039	41.721	0.000	1.561	1.715

Ljung-Box (L1) (Q):	0.42	Jarque-Bera (JB):	2067.33
Prob(Q):	0.52	Prob(JB):	0.00
Heteroskedasticity (H):	1.14	Skew:	-0.76
Prob(H) (two-sided):	0.31	Kurtosis:	11.33

Fig. 9. ARIMA model parameters

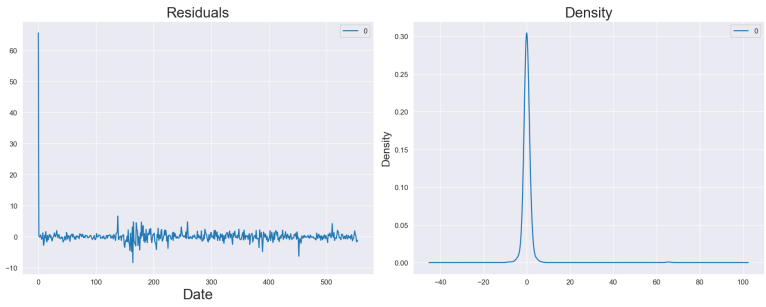

Fig. 10. ARIMA residual plot

The above plot is residual plot obtained for the ARIMA model.

It should be noted that the values of the parameters obtained are for when the model is trained taking the entire time series data into account. To evaluate how well this model works, we train a (1,1,1) model on 80 percent of the data iteratively, and compare the acquired forecast with the remaining 20 percent of the data. The results of this forecasting are shown in the following figure(Fig 11.).

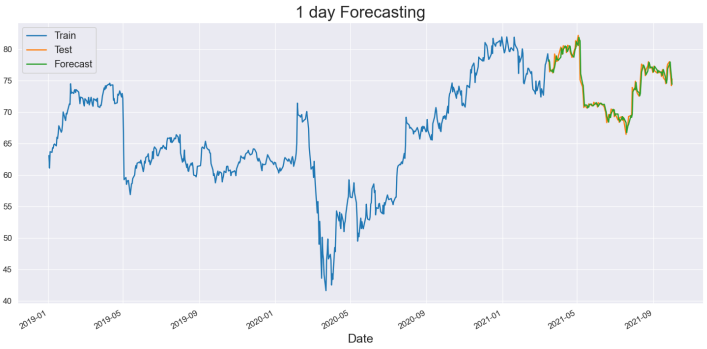

Fig. 11. ARIMA forecasting

Qualitatively, our forecast of the test data seems to very closely match the original test data set. Quantitative metrics help confirm these observations.

- **Mean Absolute Error**: 0.6809643513077142
- **Mean Squared Error**: 1.0058951127912743
- **R-squared Score**: 0.9380363414138985

These metrics obviously imply a great fit!

Next, we build an LSTM network to model the cognizant closing prices. The neural network is designed using the **TensorFlow** framework.

The network architecture includes two LSTM layers with 50 units each joined sequentially. Following this is an ordinary dense layer that outputs the prediction.

The optimizer used is Adam optimizer, and the cost function employed is the mean squared error. The network is trained for 20 epochs using mini batch gradient descent. The size of the batch is 32.

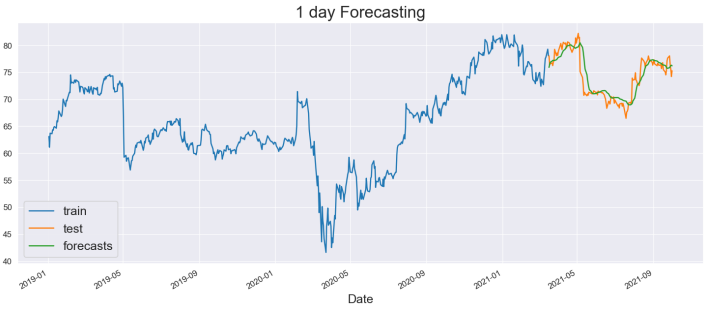

Fig. 12. LSTM forecasting

The quantitative metrics obtained for this forecast are as follows:

- **Mean Absolute Error**: 1.2927636677966226
- **Mean Squared Error**: 3.5764816453031116
- **R-squared Score**: 0.7272286305206187

These metrics obviously imply a great fit though not as great as the fit given by the ARIMA model. One reason for this may

be is that the ARIMA model was retrained every time a new data point was observed with this new point being incorporated into the new training set. This couldn't be done for the LSTM network since such neural networks take a notoriously long time to train. However, in spite of this the LSTM network still performs decently well. Therefore, the ARIMA model is perfect for making short-term predictions while an LSTM would be more apt for longer-term predictions.

B. Analyzing correlations between the prices of the different companies

Our next piece of analysis is concerning any correlations that may exist between the performances of the different companies whose stock price data has been provided. The closing price of the stock is used as a metric for stock performance.

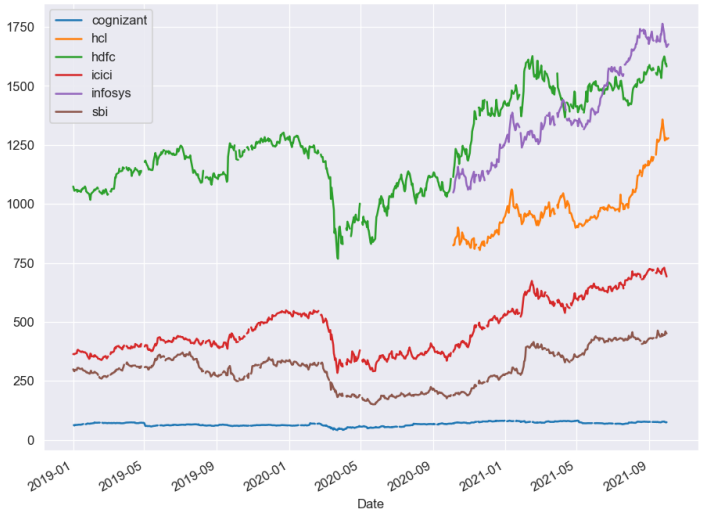

Fig. 13. Closing Prices

From the above figure, we can easily ascertain the following pieces of information: for HCL and Infosys, time series data for only the 2020-21 period is provided, and there are very few missing values present in this 2020-21 period for which there is data on all companies. Therefore, instead of imputing any rows that have null values, such rows are dropped from the table.

As a result, the following plot is obtained,

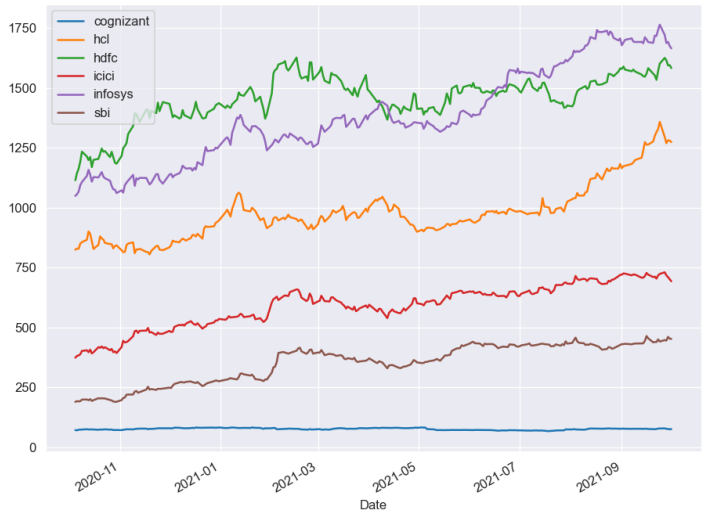

Fig. 14. Closing Prices

First, we generate the correlation matrix for this data to gain some initial understanding of the relationships between the performances of different companies.

Fig. 15. Correlation matrix between the closing prices

Looking at the values in this matrix, it is very clear that some strong linear relationships do exist. This fact is corroborated by examining the eigenvalues in the dataset.

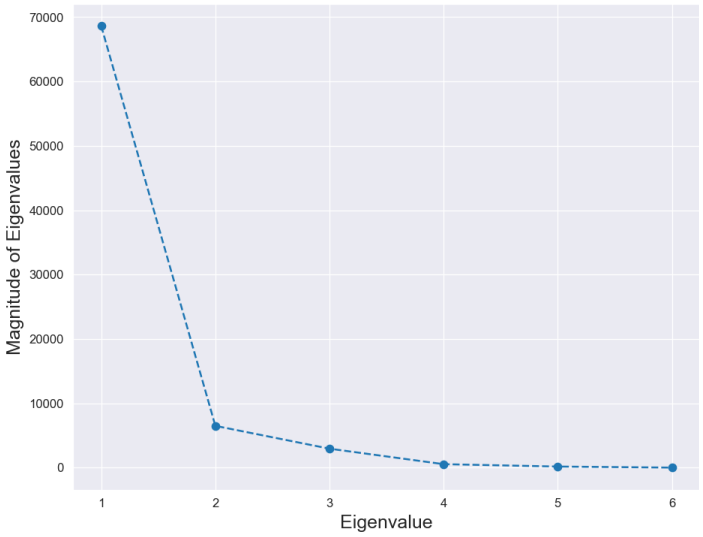

Fig. 16. Eigenvalue Plot

	EigenValues
1	68624.207267
2	6492.349221
3	2944.789643
4	540.404163
5	170.370751
6	7.123407

Fig. 17. Eigenvalues

Looking at the values of the eigenvalues we can easily say there must be at most 1 linear relation between the closing prices of the different companies: it is the eigenvector corresponding to the smallest eigenvalue.

This vector is : [0.99401884 -0.01366898 -0.02233338 -0.04944747 0.00568779 0.09361369]. These values represent the coefficients of the mean shifted variable. The first variable corresponds to Cognizant, the second to HCL, the third to HDFC, the fourth to ICICI, the fifth to Infosys, and the last to SBI.

Linear regression is also performed to get an additional interpretation of the linear relationship among the variables. For this, the dependent variable is taken to be the closing price of the HDFC stock.

The linear coefficients are shown in the following bar-plot.

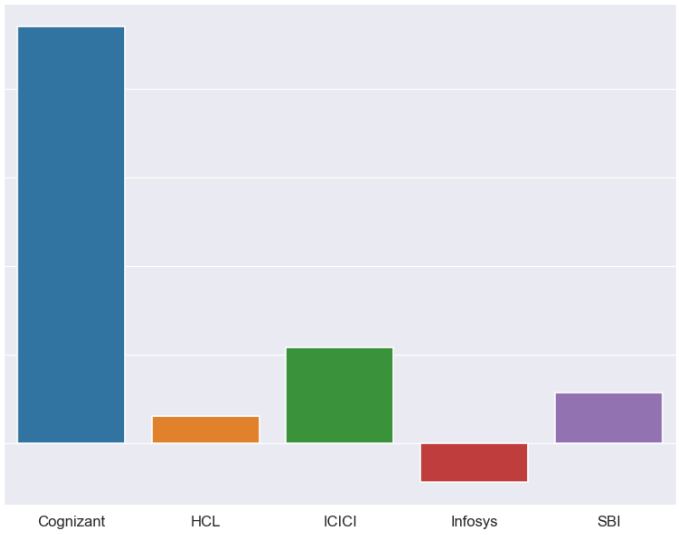

Fig. 18. Linear Regression Coefficients

The reason Cognizant has a such a high coefficient is because its closing prices are on a much smaller scale than the other companies.

The performance metrics of the Linear Regression model are:

- **Mean Absolute Error**: 26.35338962159086
- **Mean Squared Error**: 1016.7004946318239
- **R-squared Score**: 0.935600987699402

C. Analyzing correlations between the prices of the different companies and the exchange rate

The following figure is a plot of the mean shifted closing prices of the US-INR exchange rate.

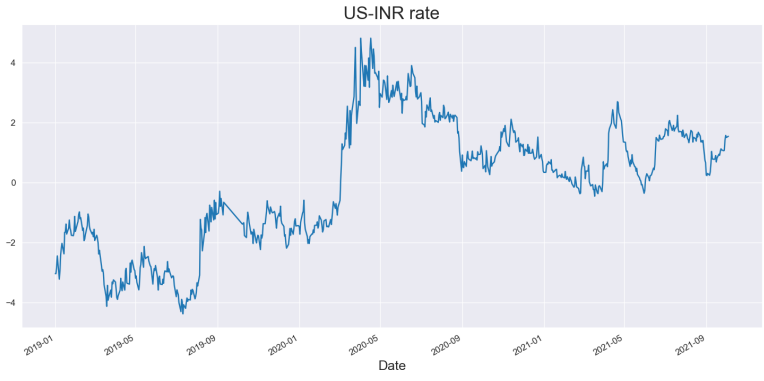

Fig. 19. US-INR exchange rate

To gain some basic understanding of the relations between the closing prices of the various companies, and the exchange rate another correlation matrix like earlier is plotted but this time the close prices of the exchange rate are also included.

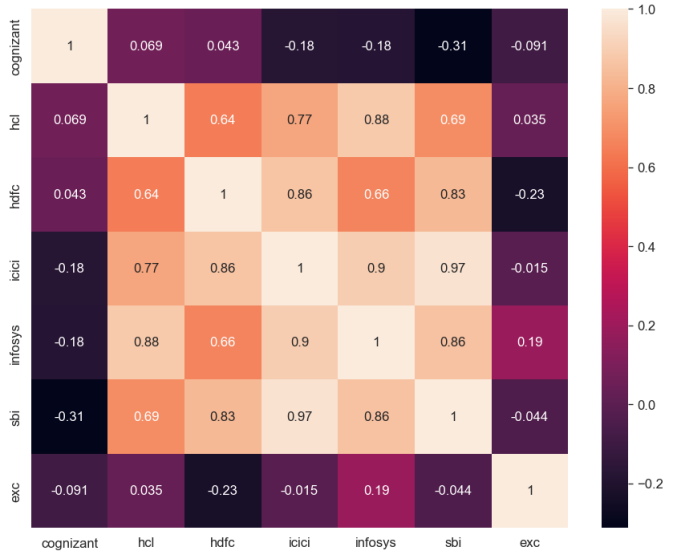

Fig. 20. Correlation Matrix

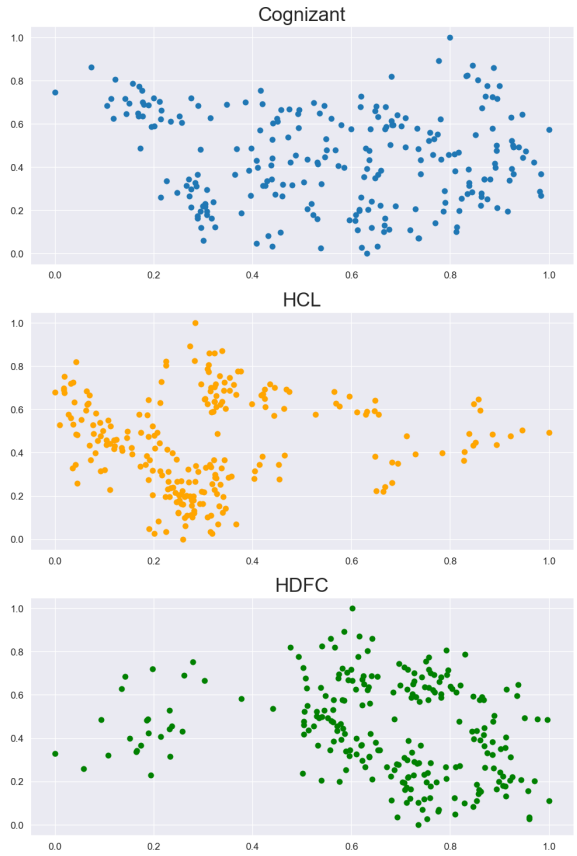

Fig. 21. scaled scatter plot of the exchange rate prices vs company stock price

The first and perhaps the most important thing this correlation matrix tells us is that the closing prices of the companies are barely linearly correlated with the target variable which is the exchange rate. This statement is supported by Fig 21.

Fig 21. is a scatter plot of the scaled company stock price vs the scaled closing price of the exchange.

To model the dependency of the exchange rate on the stock prices, we use two ensemble learning techniques: the random forest regressor and the extreme gradient boosting regressor.

- Random Forest Regressor
  - **Mean Absolute Error**: 0.24658827250001045
  - **Mean Squared Error**: 0.10251804307897268
  - **R-squared Score**: 0.8180608438009728
- Extreme Gradient Boosting Regressor
  - **Mean Absolute Error**: 0.30117832132975114
  - **Mean Squared Error**: 0.1351563656757908
  - **R-squared Score**: 0.7601374900705239

Next, we have the feature importances plots for both the models.

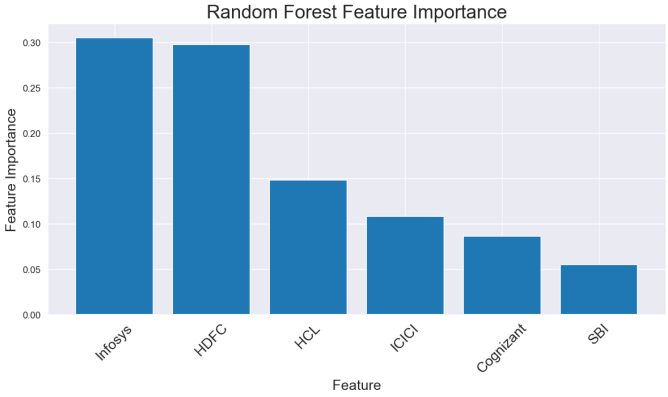

Fig. 22. Random Forest feature importance plot

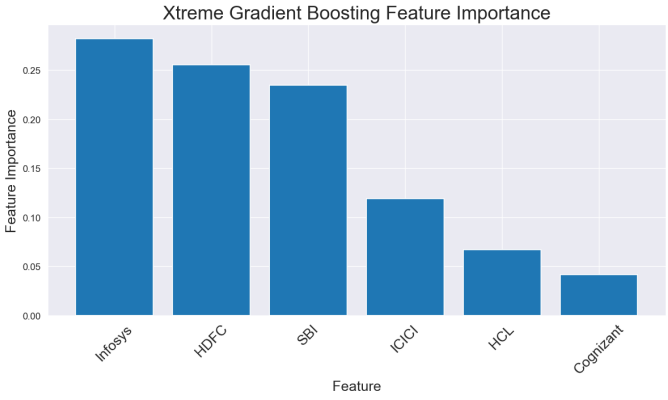

Fig. 23. Extreme Gradient Boosting importance plot

What's interesting here is that while both models assign the most significance to Infosys, and HDFC, they severly disagree when it comes to the importance of SBI. The random forest thinks it's the least important while the gradient boosting algorithm thinks its the most important feature after Infosys and HDFC.

The following figure is a plot of the scaled exchange rate prices vs the scaled SBI stock prices

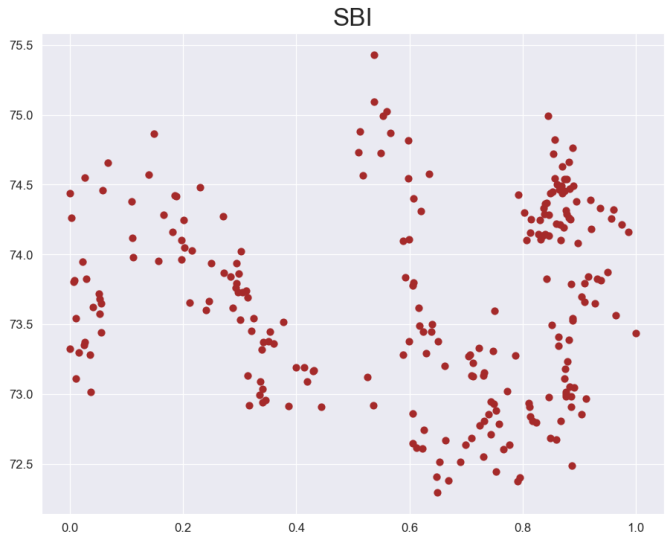

Fig. 24. Exchange rate vs SBI stock

D. Conclusions

In this analysis, we explored the intricacies of stock data for several prominent companies and the US-INR exchange rate. Utilizing a combination of traditional time series forecasting methods and advanced machine learning techniques, we aimed to uncover patterns, relationships, and potential insights within the financial datasets.

Our focus on Cognizant allowed us to showcase the effectiveness of the ARIMA (AutoRegressive Integrated Moving Average) model in forecasting stock prices. The ARIMA model, with parameters (1, 1, 1), demonstrated impressive accuracy both qualitatively and quantitatively thus making it a suitable choice for short-term predictions.

On the other hand, the Long Short-Term Memory (LSTM) neural network, a powerful tool in the realm of deep learning, presented competitive results. Despite computational challenges and the inability to retrain the model with every new data point, the LSTM exhibited great forecasting capabilities, making it particularly suitable for longer-term predictions.

Furthermore, we explored correlations between the closing prices of different companies. Eigenvalue analysis and linear regression coefficients provided valuable insights into the relationships among these companies, emphasizing the interplay of various factors in the stock market.

Ensemble learning techniques, specifically the Random Forest Regressor and Extreme Gradient Boosting Regressor, were employed to model the dependency of the exchange rate on stock prices. While both models performed well, slight discrepancies in feature importance highlighted the complexity of these relationships.

In conclusion, this analysis offers a multi-approach study on stock market dynamics. From the accuracy of forecasting models to the intricate relationships between different financial instruments, our exploration provides a foundation for informed decision-making in the volatile landscape of financial markets. As we acknowledge the limitations and assumptions inherent in our models, this study lays the groundwork for future research and improvements to further enhance our understanding of stock market behavior.

References

[1] ”Autocorrelation Plots: Graphical Technique for Statistical Data”, www.dummies.com/article/technology/information-technology/data-science/big-data/autocorrelation-plots-graphical-technique-for-statistical-data-141241/.

[2] ”ARIMA Model for Time Series Forecasting”, www.kaggle.com/code/prashant111/arima-model-for-time-series-forecasting .

[3] ”Stock Analysis: Different Methods for Evaluating Stocks”, www.investopedia.com/terms/s/stock-analysis .

[4] ”Understanding LSTM Networks”, colah.github.io/posts/2015-08-Understanding-LSTMs .