Final Exam: Seven Stock Market, Data Lab, 2022

Mustefa Abrahim
Deprt. Data Science
Indian Institute of Technology Madras (IITM)
ge22m014@smail.iitm.ac.in

***Abstract*—In this study, we predict the seven stock markets based on the given data using various methods we learned as well as new models.**
The given dataset contains seven entries, each with an opening price, a close price, and so on. Based on their characteristics, we predict the future opening price using various methods. [1]

Keywords: Coginative, HCL, Visualization, stock market, modeling

A. *Introduction*

The datasets provided consist of the time series trend of Opening price, closing price, Volume, etc. for multiple stocks such as - HCL, HDFC, ICICI, Infosys, SBI, USD vs INR exchange rate, Cognizant. [3] Filling missing values with the mean or median may introduce furher noise. Dropping the data point may cause the sampling rate to become inconsistent. Therefore, to fill in missing values the value of the previous timestamp is propagated forward using the fill method in pandas.

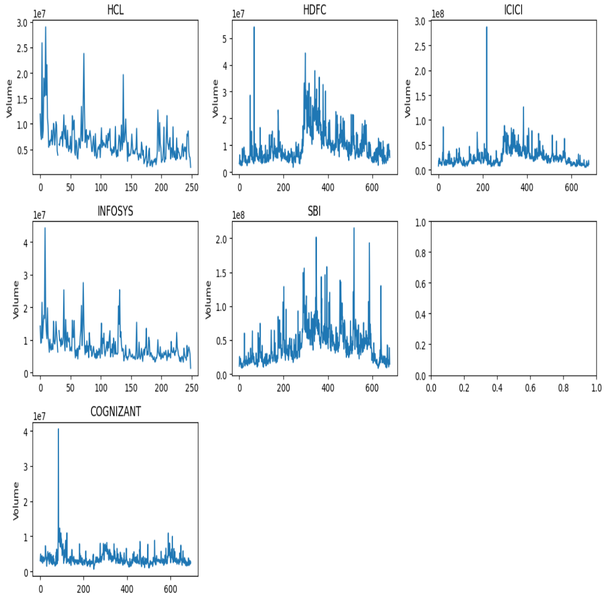
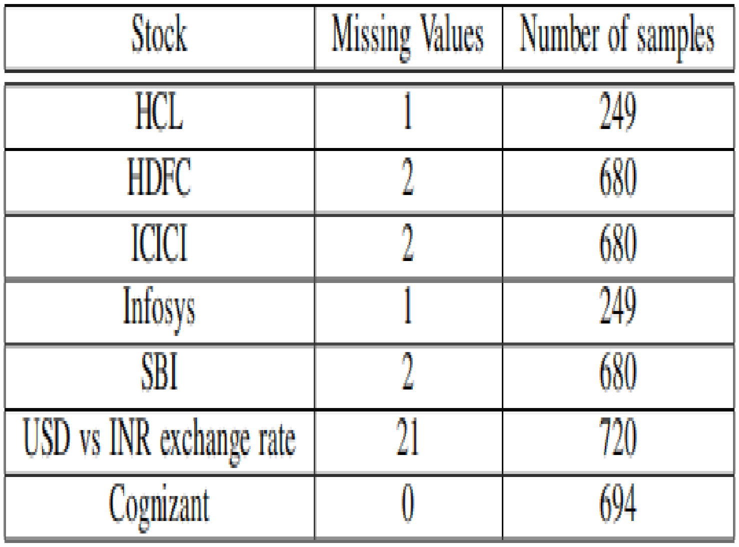


Figure 1: Visualization of datasets

I. Exploration and Visualization

we look at the trend among the closing prices of the various stocks. In HCL and Infosys we see a mostly increasing trend while in HDFC, SBI and Cognizant we see a sharp dip followed by an increase. We aim to predict the closing price for future timestamps in this study. [2] we plot the volume of stocks traded over the timestamps. We observe spikes on certain timestamps, when the volume traded was high, however the average trend remains almost constant for most stocks. No volume exists we take the moving average of the previous 10, 20, 50 days. This gives us an idea of the average trend over the past days and it performs a form of smoothing. for each stock we plot the percentage change in the closing prices daily. This is called the daily return. In HCL and Infosys again we see daily jumps more frequently, whereas in the other stocks we see a few spikes clustered in between, the trend on other days being mostly close to the average. This can further be confirmed from where we can see that the daily percentage distributions for HCL and Infosys have higher variance than the others. Following this, we merge closing prices of all stocks on the same dates. Following this we take the percentage change for each stock wrt corresponding timestamps and explore the trend between various stocks on the same date. It is seen, from that HCL shows a linear trend with Infosys, similar to the one we had previously noted Similarly, row 2 shows that on days when HDFC rises, SBI and ICICI will almost certainly rise as well. This is consistent with our intuition,



Stock	Missing Values	Number of samples
HCL	1	249
HDFC	2	680
ICICI	2	680
Infosys	1	249
SBI	2	680
USD vs INR exchange rate	21	720
Cognizant	0	694

Figure 2: Missing values of the given dataset

as these are all banking stocks. Cognizant and USD vs INR do not exhibit a significant trend when compared to any other stock. [1] The correlation coefficients for each of these relationships are given. We create a PairGrid using the Daily Percentage data from all of the stocks. The diagonal is made up of the value histogram. The upper triangle is a scatterplot, while the lower triangle is a plot. [2] We can see from the plot that there are multiple plots of daily returns of the stocks. Methods: A. Techniques learnt in the course:

- (i) Linear Regression
- (ii) Logistic Regression
- (iii) Random Forest

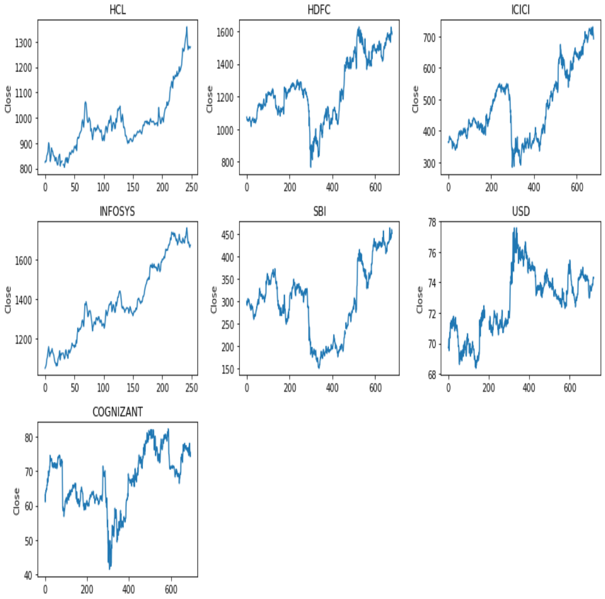


figure 3: visualization plot

II. Data

Data is available everywhere, and obtaining data is not difficult. We have seven stock market datasets, including the USD-INR exchange rate 2019_2021, Cognizant share prices 2019_2021, and HCL Technologies share prices 2019 2021. SBI Share Prices 2019_2021, HDFC Bank Share Prices 2019_2021, ICICI Bank Share Prices 2019_2021, Infosys Share Prices 2019_2021, and HDFC Bank Share Prices 2019 2021. And we predicted the stock market using various methods we learned in class, such as linear regression, logistic regression, decision trees, random forests, and so on. [3]

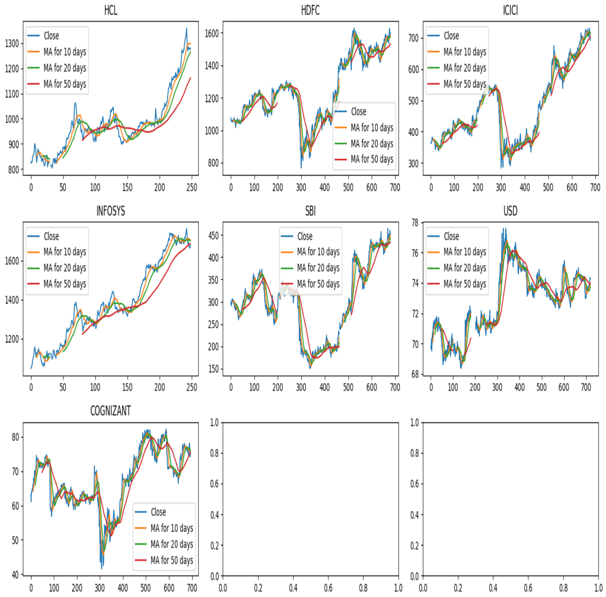


Figure 4: visualization 2

III. Modeling

Data modelling is the process of visualising an entire information system or some of its components in order to communicate the relationships between data points and structures. The goal is to demonstrate the different types of data stored in the system, the relationships between them, the formats and attributes of the data, and how the data can be grouped and organised. [1]

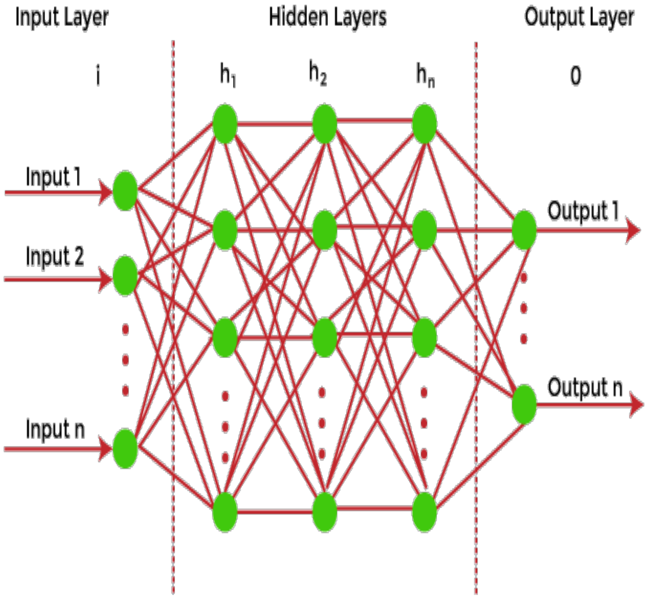


figure 5: Machine learning models

Data Models are typically developed in response to business requirements. Requirements and rules are defined upfront using feedback from business stakeholders in order to be used in designing a new system. The Data Modeling process starts with gathering information about business requirements from stakeholders and end-users.

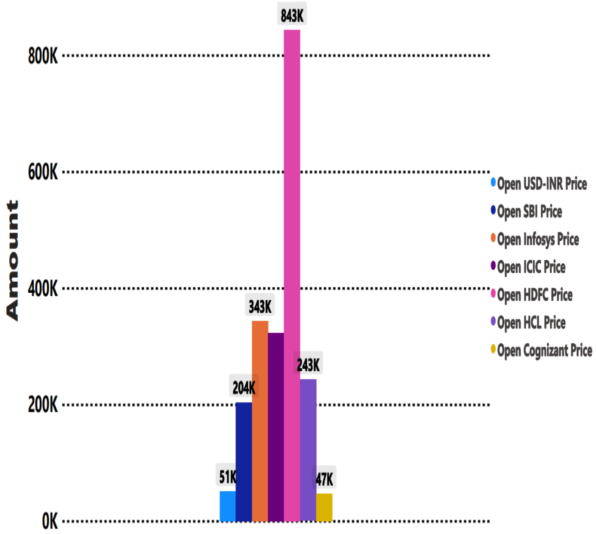


Figure 6: Open price

Here is the visualization the seven stock markets the opening and closing change of each market.

We discovered differences in the high change from one bank to another as well as the low change from branch to branch. Additionally, the date of these exchanges gave us an overview of the exchange at this time of year, as opposed to different days and months, which we found. There were seven data sets, each of which had a different number of instances. They were all the same with different banking branches.

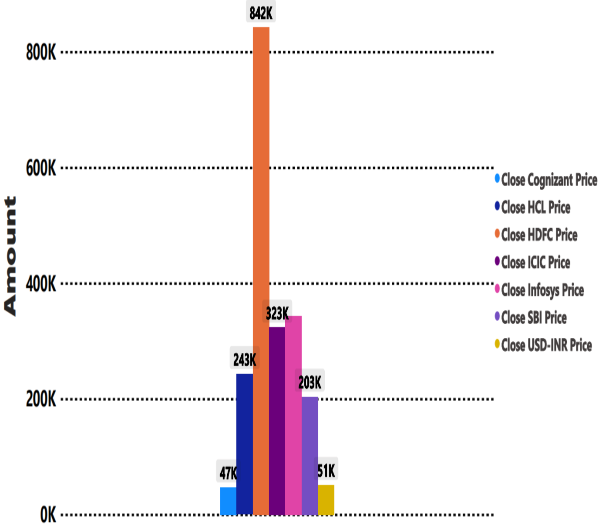


Figure 7: close change

IV. Conclusions

In this paper, we predict the share of seven stock market data-sets , from the Analysis to the modelling. And We discovered that the linear regression model is more powerful than other models for predicting stock market movements than other models that we learned. And by fine-tuning some of its parameters, we were able to avoid the problem of over-fitting and achieve over 80% accuracy.

References

[1] Hastie T.J. Springer; New York: 2008. The elements of statistical learning: data mining, inference, and prediction. [Google Scholar]

[2] Amit, Y., Blanchard, G., Wilder, K. (1999). Multiple randomized classifiers: MRCL Technical Report, Department of Statistics, University of Chicago.

[3] Boccaletti S. Modeling and forecasting of epidemic spreading: the case of COVID-19 and beyond. Chaos Solitons Fractals. 2020 [PMC free article] [PubMed] [Google Scholar]