Final-Exam: Analysis of Share Prices

Shania Mitra
Roll Number: CH18B067
Chemical Engineering
IIT Madras, Chennai
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we revisit all previous assignments and present them in a succinct manner. Further we analyse data from 7 stocks and make predictions on the closing price for future timestamps.**

***Index Terms*—Linear Regression, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Classifier, Time Series, ARIMA**

I. Introduction

This essay is divided into two main parts. In the first part we try to present the previous mathematical models on various models in a more crisp and concise manner, presenting findings in a more orderly fashion. This part consists of 6 sections. In the following part, we present the analysis of share prices using methods already learnt and new methods along with detailed exploratory analysis on the time series of each stock.

II. Linear Regression

Linear Regression is an linear approach for modelling the relationship between a scalar response and one or more explanatory variables, by taking a linear combination of input variables. In this study, we use linear regression to explore the correlation between cancer incidence and mortality rates and socio-economic and racial factors. We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. The aim here was to examine whether low income groups are at a greater risk of succumbing to cancer and to to demonstrate whether cancer incidence is correlated with socio-economic status. To analyse the dataset, new features were created:

- Total Number of Males: Obtained by summing number of males insured and number of males not insured, similarly we obtain the total number of females (It was verified that number of males with insurance + number of males without insurance + number of females with insurance + number of females without insurance == all with insurance + all without insurance)
- Total Population: The total population of the area is obtained by summing number of males and females
- Female Ratio: Number of females divided by the total number of people, similarly we can obtain ratio of males in the area
- Female Poverty Ratio: Ratio of poor females among total number of females, similarly Male Poverty Ratio can be obtained
- Poverty Ratio: Number of poor people divided by total number of people in that area
- Female Insurance Ratio: Number of females insured divided by total number of females, similarly, we can obtain male insurance ratio
- Number of Groups above Median: This is a number ranging from 0 to 5 indicating the number of ethnic groups having median income above the median income of the entire population

Upon analysing the data we make the following observations:

- As the median income increases mortality rates decrease
- As the rate of incidence increases, mortality rates also increase
- Blacks have the lowest mean 'median income' while asians have the highest mean 'median income'
- Most blacks and hispanics are payed low incomes hinting at racial discrimination
- In regions with a falling trend in cases, the median "median income" is higher than the median in other regions, suggesting that better income may help people afford better healthcare and facilities, leading to a falling trend in cases.
- Areas with more ethnic groups earning above the median salary are more likely to subscribe to an insurance. This insurance subscription, does not, however, translate to a lower incidence rate or a lower mortality rate
- Among all ethnic groups, subgroups with higher median income tend to show a falling trend in cases
- On an average females are more poverty-stricken than males
- In regions where poor females are more in number as compared to poor males, incidence and mortality rates are higher
- Females, who tend to be poorer, are more prone to cancer

After building the models it can be said that there is some correlation between socioeconomic factors and incidence/mortality rate (since, upon inclusion of those factors prediction capability of the model improves) however it is not very strong since test performance is not satisfactory in each of the cases. The results can be seen in Figure 1.

III. Logistic Regression

Logistic Regression is a supervised machine learning algorithm that can be used to model the probability of a certain

OLS Regression Results

Dep. Variable:	incidence_rate	R-squared (uncentered):	0.969
Model:	OLS	Adj. R-squared (uncentered):	0.968
Method:	Least Squares	F-statistic:	1344.
Date:	Fri, 17 Sep 2021	Prob (F-statistic):	0.00
Time:	11:04:33	Log-Likelihood:	-2263.7
No. Observations:	581	AIC:	4553.
Df Residuals:	568	BIC:	4610.
Df Model:	13		
Covariance Type:	nonrobust		

Fig. 1. Results for Assignment-1: Linear Regression

class or event. It is used when the data is thought to be linearly separable and the outcome is binary or dichotomous in nature. In this study, we use logistic regression to model probability of survival of groups of people on the RMS Titanic as a function of age, socioeconomic factors, number of family members, port of embarkment etc.

To analyse the dataset, the new feature of family size was introduced.

The key insights from the study are as follows:

- Among the 891 passengers in the training set provided, 549 were found dead while 342 managed to survive
- As number of siblings/spouses increase, number of parents and children also increases.
- People who escaped also had a lower average age, which aligns with the fact that younger people due to higher agility find it easier to escape
- The ship had fewer females yet the survival rate among females is much higher
- Passengers in higher PClass had lower survival rates and higher Pclasses have lower fares. Thus, passengers who paid higher fares were more likely to survive. This may have happened due to more expensive cabins being located at more strategic locations in the ship, allowing the people in them to escape more easily
- In the 1st Class, the survival rates are much higher than death rates. For 2nd Class, they are almost the same whereas for the 3rd Class it is much lower. This may also be attributed to the fact that the number of people in 3rd Class was much higher than those in 1st or 2nd Class, making it difficult for people to escape from the stampede
- Most of the passengers on the ship were individuals. However, the survival rate among individuals was very low.
- In contrast, the survival among families with 2, 3 and 4 people was higher than death rates, possibly due to members among the family helping eachother to escape. In large families, however, death rates were very high possibly because of family members not wanting to escape without eachother and the lack of lifeboats restraining their escape
- Most of the people on the ship were from port S, however, survival rates were highest among passengers from Port C.

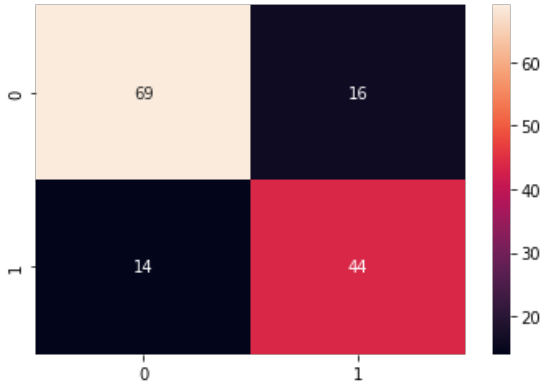

Fig. 2. Results for Assignment-2: Logistic Regression

Upon building the Logistic regression model, we achieve an f1-score of 0.72 while for the people who were unable to survive, we achieve an f1-score of 0.84. The number of correct classifications and misclassifications can be seen from the confusion matrix in Figure 2.

IV. Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of conditional independence among predictors. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. In this study, we use a Naive Bayes classifier to examine whether a person makes over $50K a year based on their age, educational qualification, marital status, occupation, race and other factors. Upon analysing the data we make the following observations:

- No individuals having gone through only preschool receive wages above $50K. For all other degrees, individuals receiving both below and above $50K exist.
- Among those who have never worked or are working without pay, only whites and blacks are found
- Among those who work without pay, Whites, Blacks and Asians can be found

- Amer-Indian Eskimos and Blacks among others have the highest proportion of people with wages below $50K annually.
- Those who have never worked belong to junior and middle school, while those who work without pay mostly belong to middle and high school.
- At lower ages (17-25) the average working hours are lower than 40hrs and increases with increasing age. From 25- 61, the peak working age, working hours plateau at around 45 hours/week. Beyond 61, working hours decrease with increasing age, as individuals get older and their working capacity reduces
- Males have higher average working hours per week as compared to females.
- Maximum people in the population work in the private sector. Further, for all classes except those self employed, more people earn below $50K than above.
- The mean working hours for males is larger than that of females.
- The median number of years of education for both males and females is the same
- The mean working hours as well as number of years of education for wage class >$50K is larger than that of wage class <$50K
- The median age of those who earn above $50K is higher than those who earn below $50K.
- The number of males in both wage classes is higher than that of females
- The ratio of people with income above $50K to those below $50K is larger for males as compared to females, as can be seen below as well as in the figure.
- There are very few people with edu-years <=8 and the percentage of them earning more than $50K is extremely low. For edu-years > 9-13, the number of people increases but the number of people in wage class 1 is lower than that in wage class 0. For people with more than 14 years of education however, more people earn above $50K than below.
- Husbands and wives have the largest proportion of individuals earning more than $50K while those that live separately or are unmarried have the lowest proportion. This may be due to the fact that husbands and wives need to earn enough to make ends meet for themselves as well as their children, while those that are not in family need to earn enough to fend for themselves only.

Upon building the Naive Bayes model, for the people who earn below $50K, we achieve an f1-score of 0.88 while for the people who earn above $50K, we achieve an f1-score of 0.43. The number of correct classifications and misclassifications can be seen from the confusion matrix in Figure 3.

V. Decision Trees and Random Forests

A decision tree is a set of cascading questions. When we get a data point (i.e. set of features and values), we use each attribute (i.e. a value of a given feature of the data point) to answer a question. The answer to each question decides

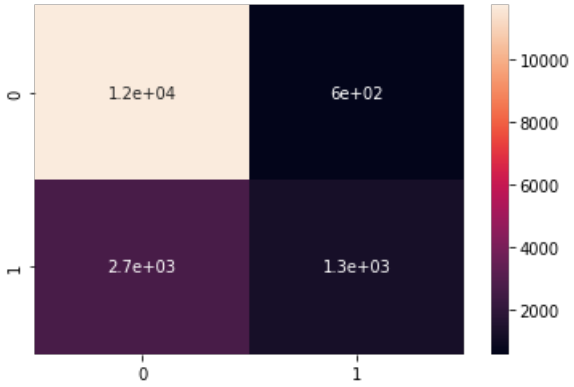

Fig. 3. Results for Assignment-3: Naive Bayes

the next question. Random forest is a supervised learning algorithm which builds an ensemble of decision trees. It is based on the principle that a combination of learning models increases the overall result. The key insights from the study are as follows:

- The class “unacc” occurs the most frequently, followed by “acc”
- If a car has very high maintenance, the target would most likely not be “vgood” or “good”
- With cars of capcity 2, only cars of target “unacc” occur. Thus, if a car has a capacity of 2, possibly a luxury or sports car, the target would most likely be “unacc”
- Cars with small boot space no samples with target “vgood” occur. Thus, if a car has small bootpsace, such as an SUV, it most likely would not have a target of “vgood”
- Cars with low safety have target as “unacc” only, while, cars with medium safety have no samples with target “vgood”. Thus, if a car has a target of vgood, it will most likely have high safety
- Cars with low safety have target as “unacc” only, while, cars with medium safety have no samples with target “vgood”. Thus, if a car has a target of vgood, it will most likely have high safety

Upon building the decision tree model, we get the highest f1 score of 0.98 for the “unacc” class while the lowest is 0.79 for “vgood”. For the random forest model, the highest f1 score remains the same while the lowest is at 0.88 for “vgood” suggesting a significant improvement. The accuracy however remains the same for both at 0.96. The number of correct classifications and misclassifications can be seen from the confusion matrix in Figure 4(a) and (b).

VI. Support Vector Machine

Support Vector Machine is a linear model for classification and regression problems. The algorithm creates a line or a hyperplane which separates the data into classes. In this study, we use Support Vector Machines to model whether a star is a pulsar or not based on statistical properties of the integrated profile and DM-SNR curve.

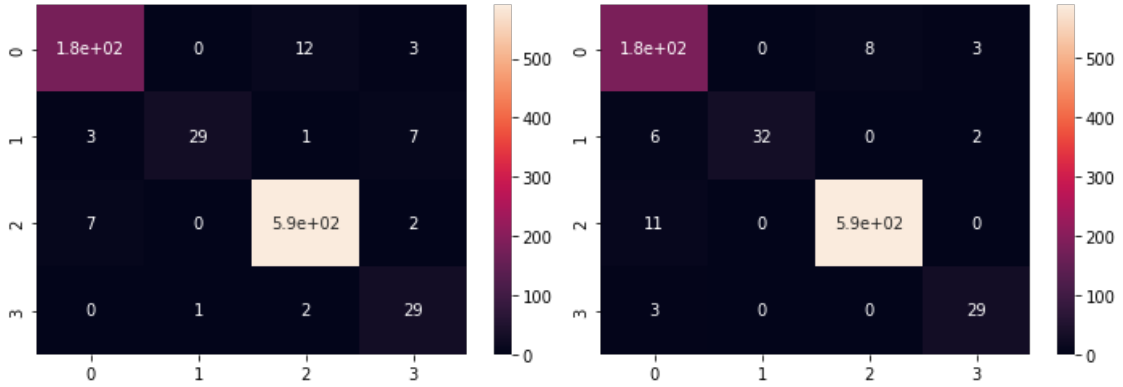

Fig. 4. Results for Assignment-4: Decision Tree and Assignment-5: Random Forests

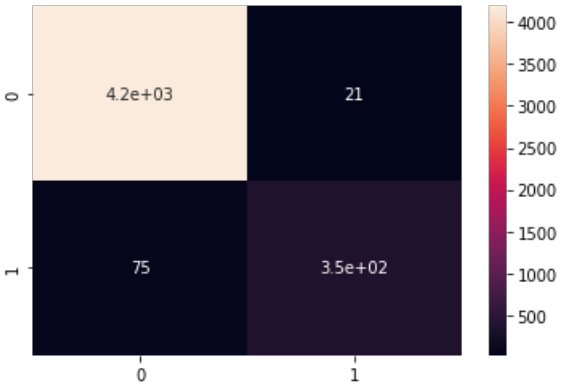

Fig. 5. Results for Assignment-6: Support vector Machine

- The distributions of "not pulsar" and "pulsar" have different means and variances. The pulsar stars tend to have a higher variance and smaller peak
- Non-pulsars tend to have low or negative kurtosis and a small skew relatively independent of eachother while pulsars tend to have high skew and kurtosis, approximately linear dependent on eachother
- The target class is highly positively correlated to the skew and kurtosis of the integrated profile and negatively correlated to the mean and standard deviation of the integrated profile

On building the support vector machine model, we obtain an f1-score 0.88 on the pulsar class and a weighted accuracy of 0.98 overall. There are only 96 misclassifications among the 4637 stars, with the model confusing some pulsars to be non-pulsars due to the imbalance in the dataset. The number of correct classifications and misclassifications can be seen from the confusion matrix in Figure 5.

VII. Final Exam: Stock Market Prediction

A. Introduction

The datasets provided consists 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.

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 missing values the value of the previous timestamp is propagated forward using the ffill method in pandas.

B. Exploration and Visualization

In figure 6, 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.

In figure 7, 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

In figure 8, 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.

In Figure 9, 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 Figure 10, 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 Fig 11, that HCL shows a linear trend with Infosys, similar to

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

TABLE I
MISSING VALUES AND NUMBER OF SAMPLES FOR EACH

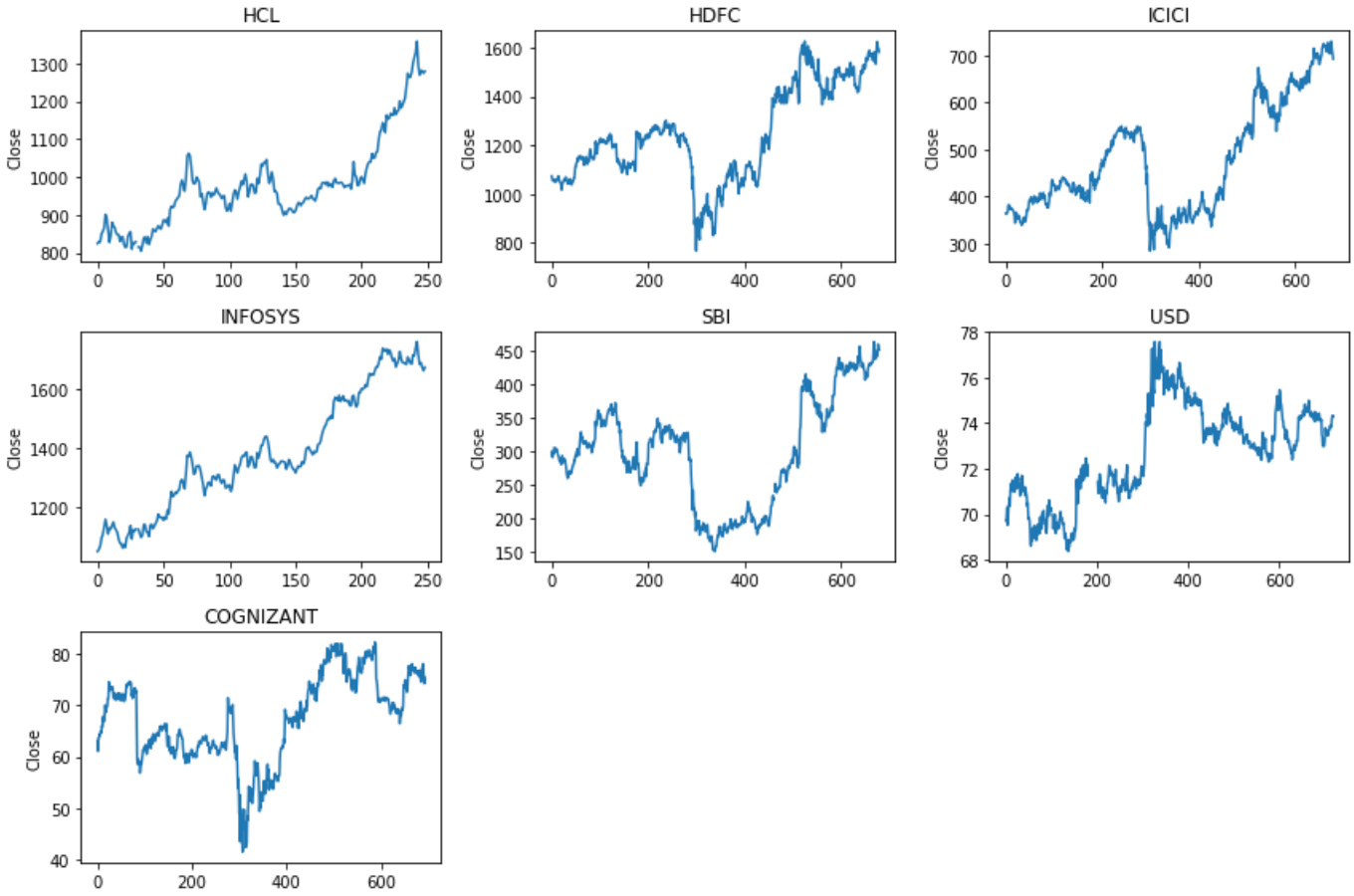

Fig. 6. Closing Price of 7 stocks plotted against time

the one we had noted before. Similarly, from row 2, it is seen that on days in which HDFC increases, it is highly likely that SBI and ICICI also increase. This matches with our intuition since each of these are banking stocks. Cognizant and USD vs INR do not show a significant trend with any other stock. The correlation values for each of these relations is given in Figure 12.

In Figure 13, we plot a PairGrid with the data of Daily Percentage of all the stocks. The diagonal consists of the histogram of values. The upper triangle consists of a scatterplot and the lower triangle consists a kdeplot. From the kdeplot we observe that in each of the plots of daily returns of the stocks, we have multiple peaks.

In Fig 14, we portray the lag plot of the daily returns of the stocks with a lag of 5. It can be seen that there is a clear linear relationship.

C. Models

In this part we discuss the models used for prediction in this study

Aim: To predict the closing price of the last 20% of the timestamps (green in Fig 6) using the training data (blue in Fig 6)

Methods:

A. Techniques learnt in the course:

(i) Linear Regression

(ii) kNN

(iii) Random Forest

B. New techniques:

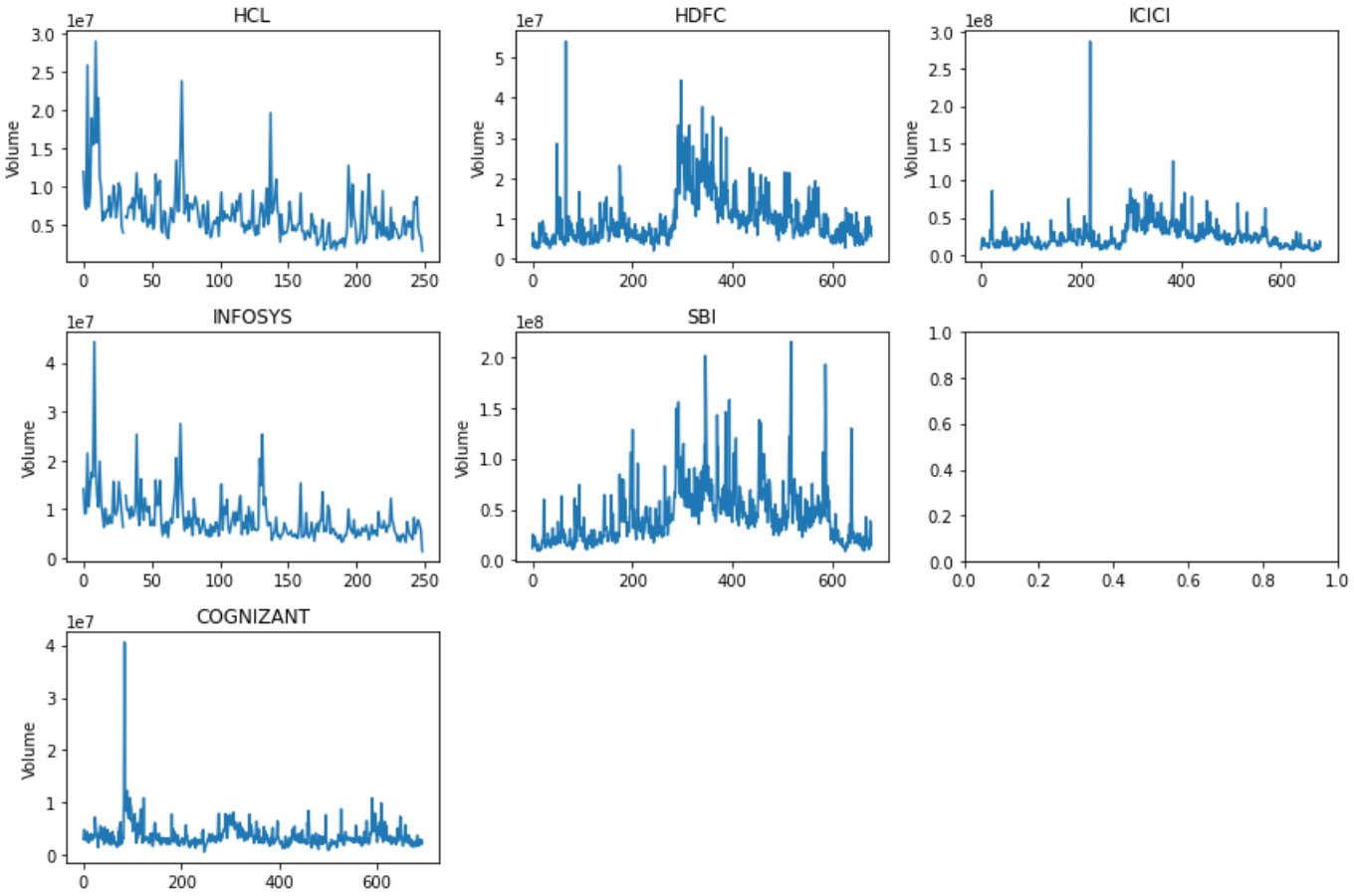

Fig. 7. Volume of 6 stocks plotted against time

(i) ARIMA

1) Techniques Learnt in the course: In this section we use Linear Regression, K-nearest neighbours and Random Forest models to predict values of future timstamps. The metric used to gauge the performance is Root Mean Squared Error, between validation samples and predictions. To use these models, th following features are used for each company:

- Opening Price
- Highest Price
- Lowest Price
- Volume Traded
- Moving average for the past 10 days (not including the current one since that would be feeding some form of the output to the model)
- Moving average for the past 20 days
- Moving average for the past 50 days
- Daily Return for the previous timestamp

REFERENCES

[1] L. (2021a, March 25). Stock Market Analysis + Prediction using LSTM. Kaggle. https://www.kaggle.com/lonnieqin/stock-market-analysis-prediction-using-lstm

[2] P. (2019, May 21). Stock Market Analysis and Time Series Prediction. Kaggle. https://www.kaggle.com/pierpaolo28/stock-market-analysis-and-time-series-prediction

[3] The Winton Stock Market Challenge | Kaggle. (2021, December 10). Https://Www.Kaggle.Com/c/the-Winton-Stock-Market-Challenge.

[4] Singh, A. (2021, July 23). Stock Price Prediction Using Machine Learning | Deep Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/

[5] Gandhi, Rohith. "Support Vector Machine — Introduction to Machine Learning Algorithms." Medium, 5 July 2018, towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[6] "Sklearn SVM (Support Vector Machines) with Python." DataCamp Community, www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python. Accessed 28 Nov. 2021.

[7] Random Forest. (2021, November 4). In Wikipedia. https://en.wikipedia.org/wiki/Random_forest

[8] Donges, N. (2021, September 17). A Complete Guide to the Random Forest Algorithm. Built In. https://builtin.com/data-science/random-forest-algorithm

[9] Z. (2021, September 26). Random Forest Explained - Towards Data Science. Medium. https://towardsdatascience.com/random-forest-explained-7eae084f3ebe

[10] Naive Bayes classifier. (2021, October 21). In Wikipedia. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[11] Zhang, Z. (2020, February 6). Naive Bayes Explained - Towards Data Science. Medium. Retrieved October 21, 2021, from https://towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0

[12] Ray, S. (2021, August 26). Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples. Analytics Vidhya. Retrieved October 21, 2021, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

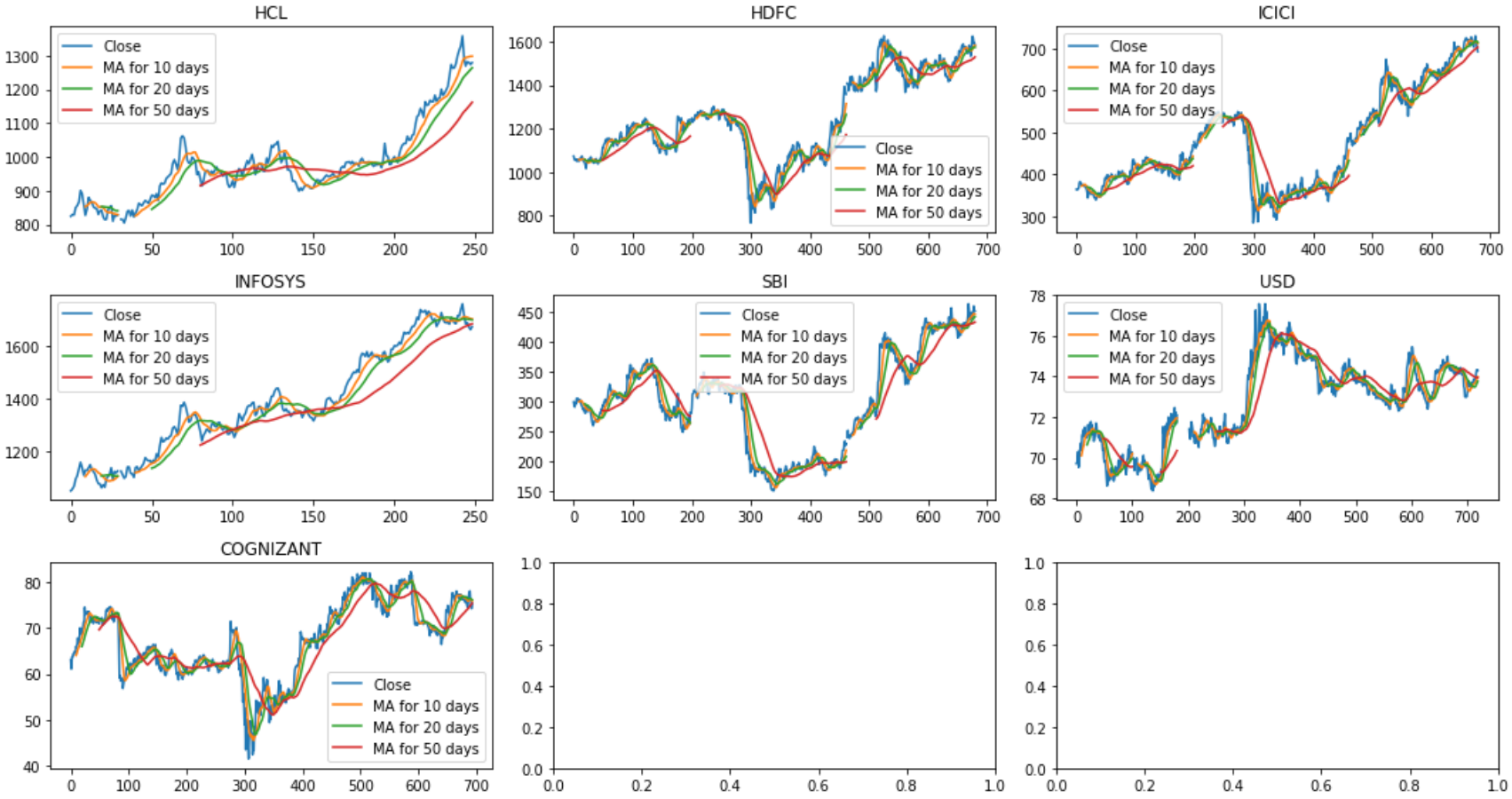

Fig. 8. Closing Price and Moving Averages of 7 stocks plotted against time

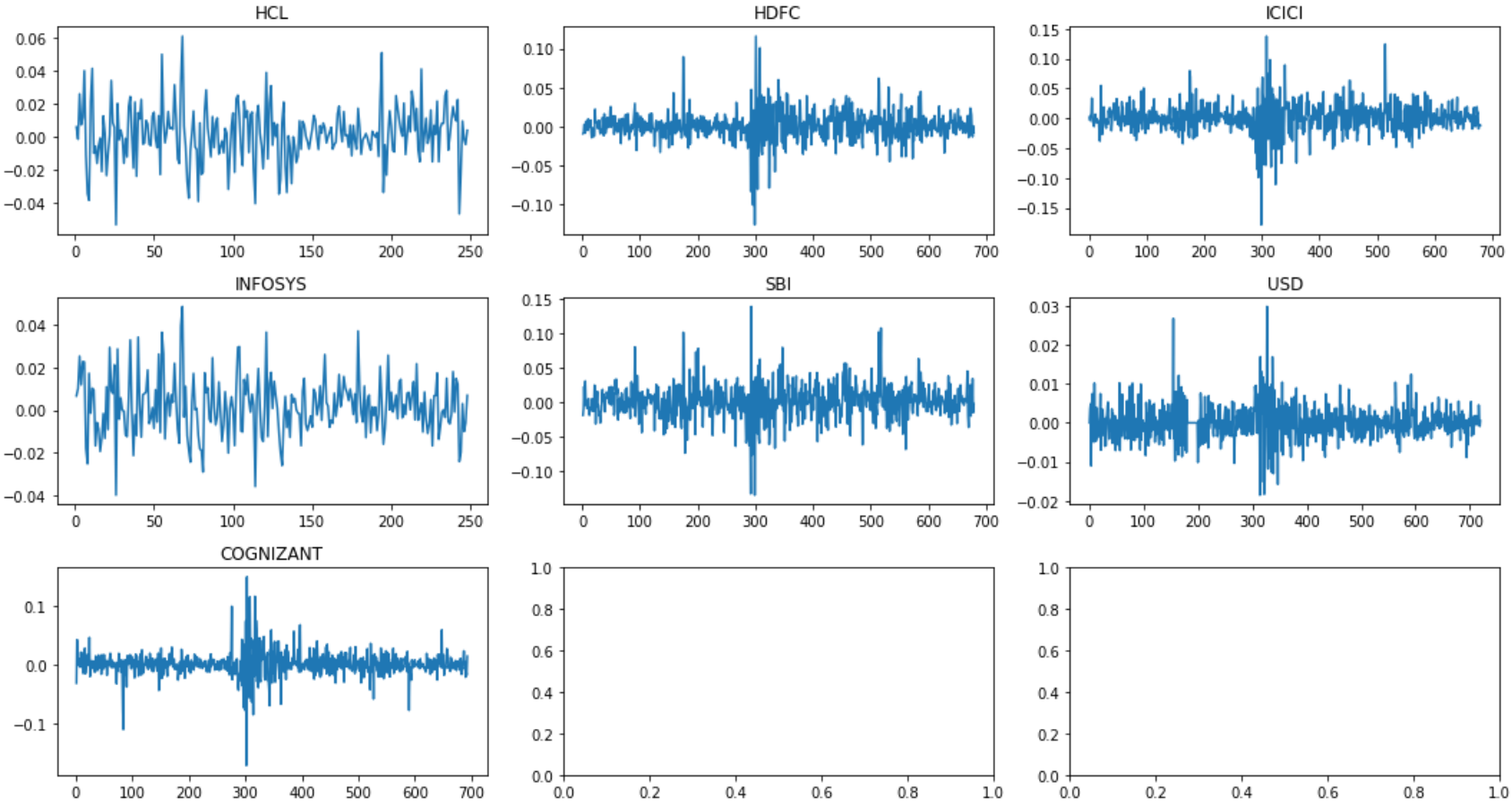

Fig. 9. Daily Return of 7 stocks plotted against time

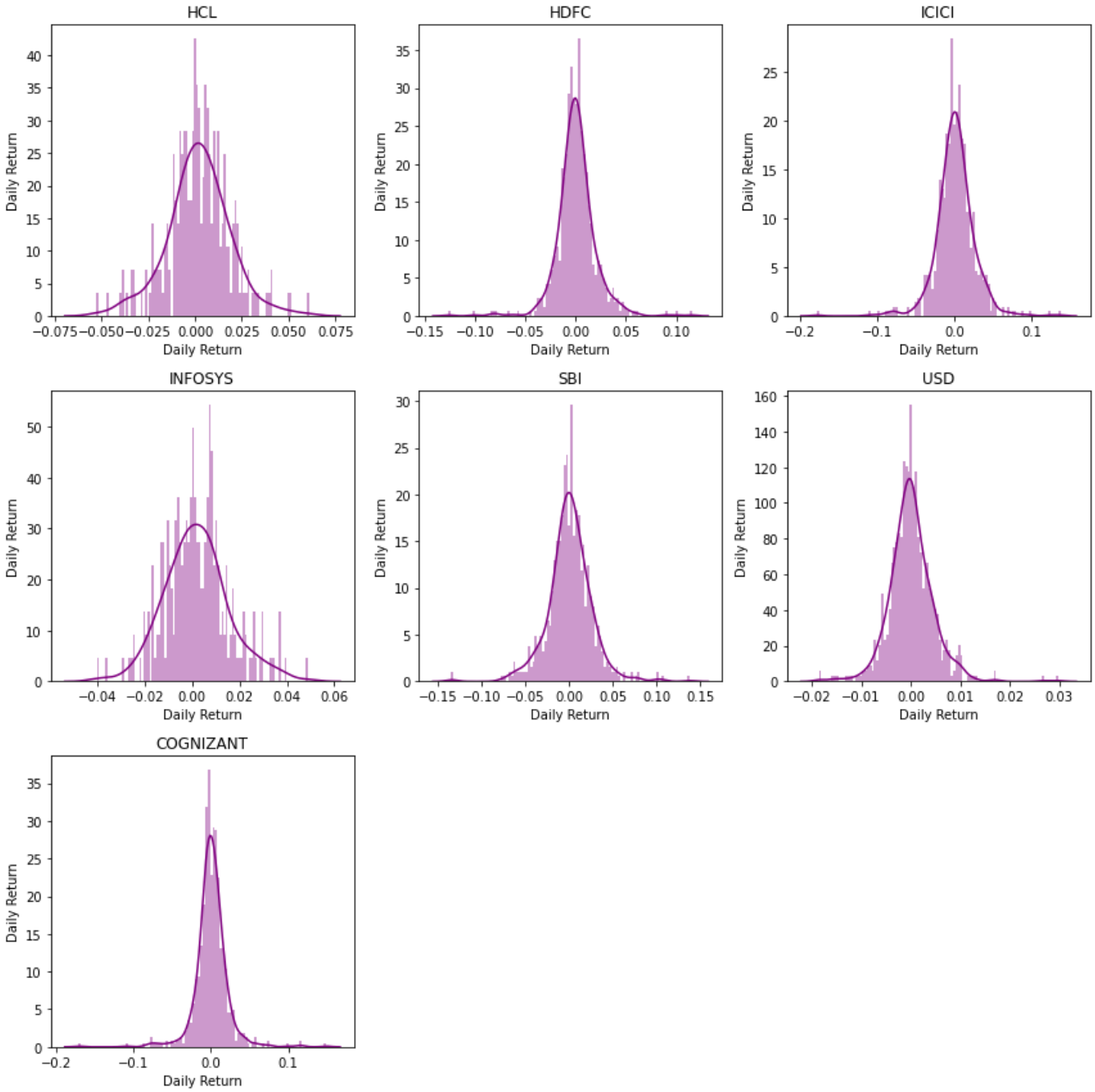

Fig. 10. Distribution of Daily Return of 7 stocks

Stock	RMSE for LR	RMSE for KNN	RMSE for RF	RMSE for ARIMA	Train Samples	Val Samples
HCL	6.522	224.539	177.575	21.241	199	50
HDFC	6.481	357.042	25.54	18.157	544	136
ICICI	3.245	189.454	65.063	9.143	544	136
Infosys	10.295	238.099	92.046	20.213	199	50
SBI	2.547	110.984	60.513	6.972	544	136
USD	0.001	1.479	0.539	0.205	576	144
Cognizant	0.318	13.27	1.547	1.019	555	139

TABLE II
RMSE FOR EACH OF THE 7 STOCKS FOR ALL METHODS

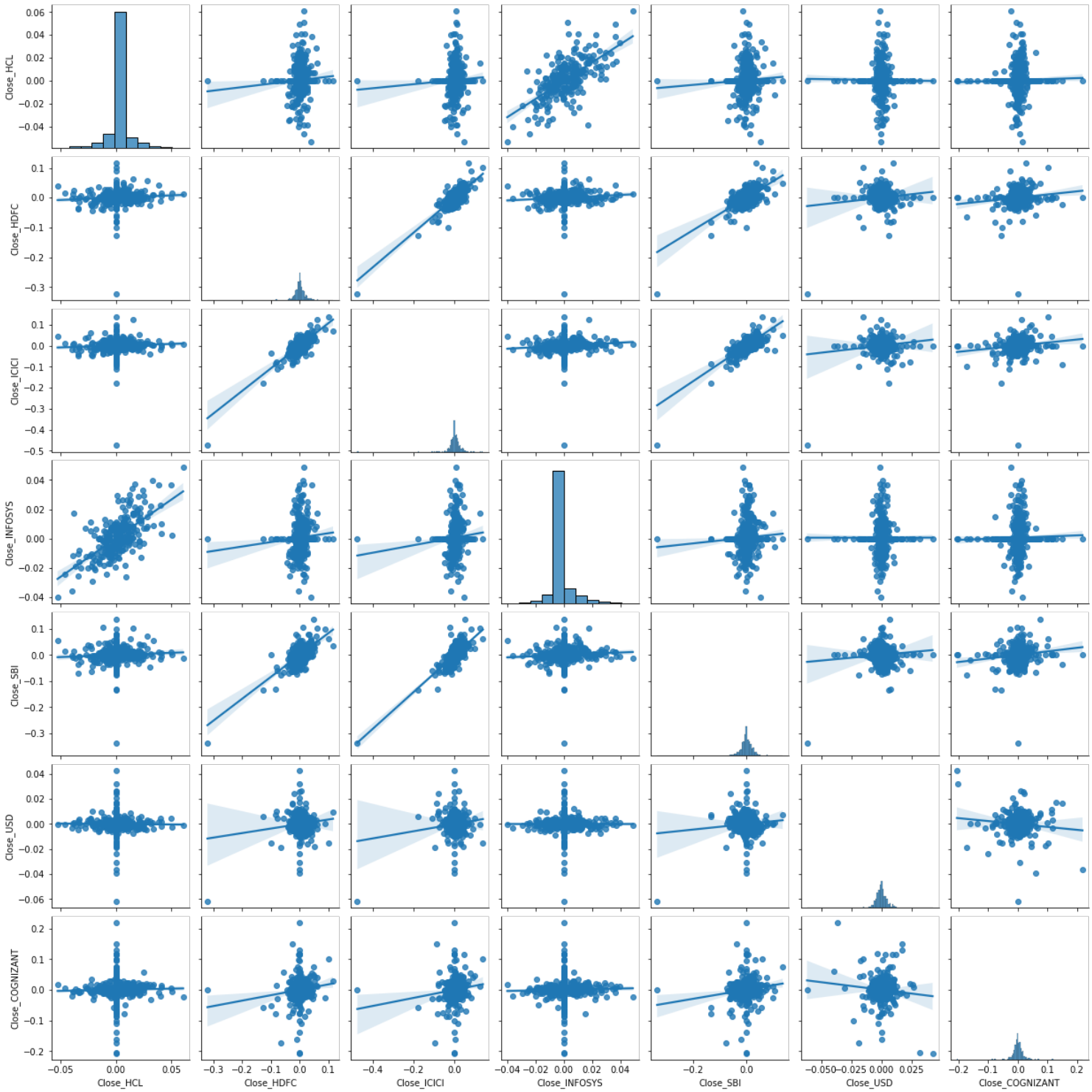

Fig. 11. Pair plot of percent changes of the 7 stocks

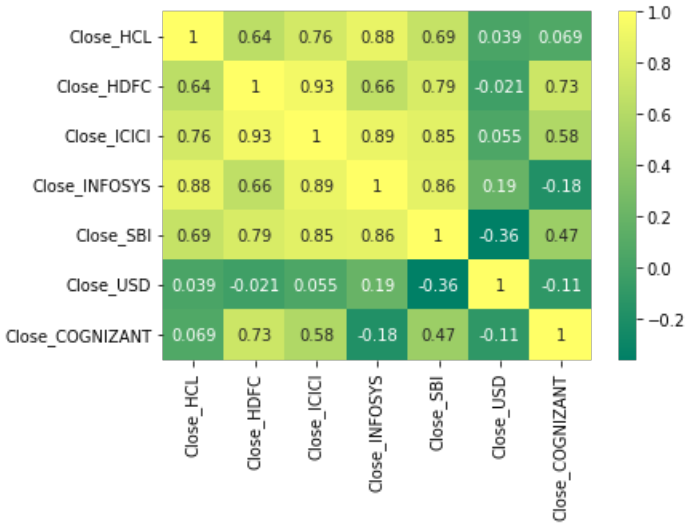

Fig. 12. Correlation plot of percent changes of the 7 stocks

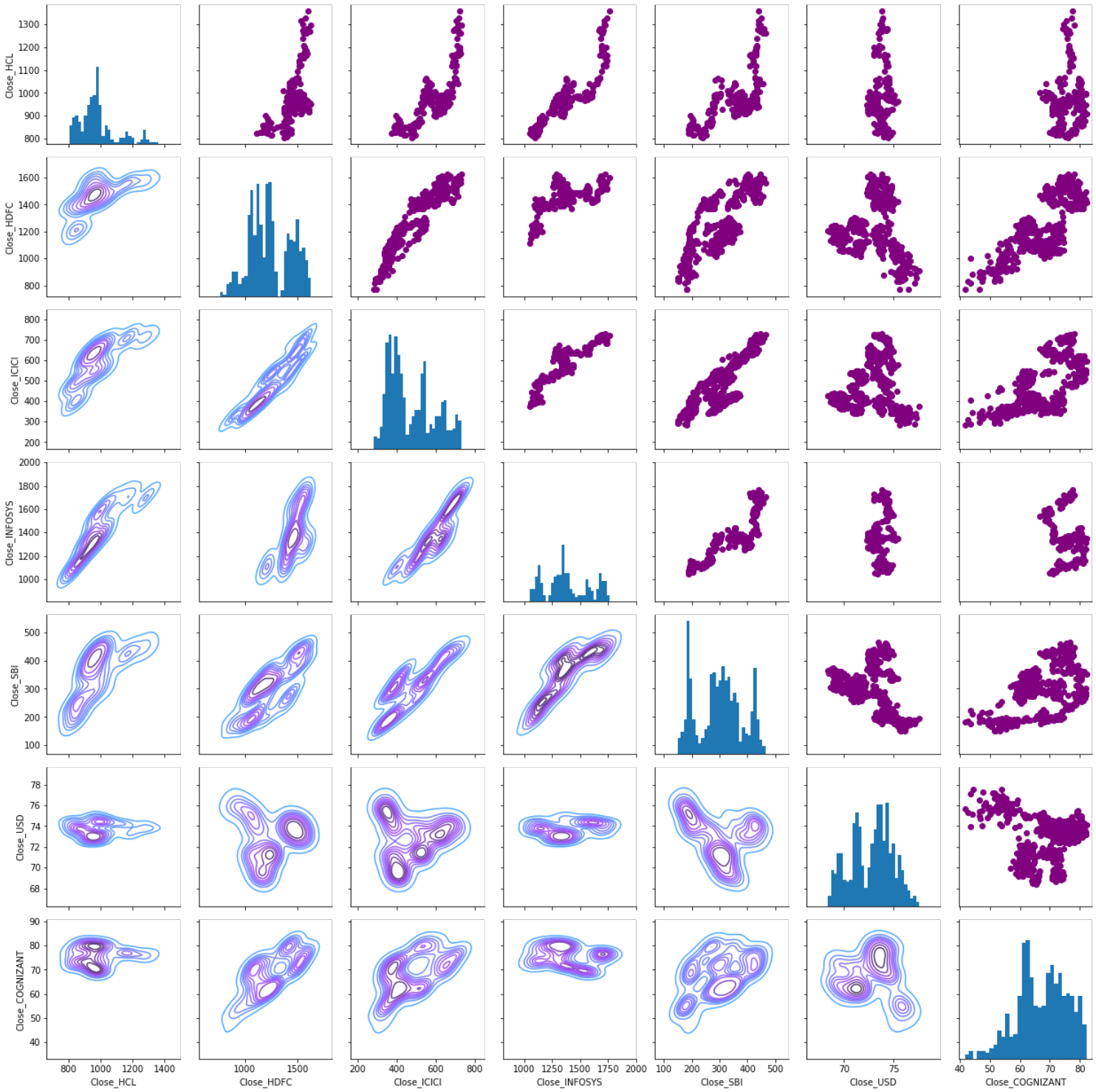

Fig. 13. PairGrid of Daily Percentage values of the 7 stocks

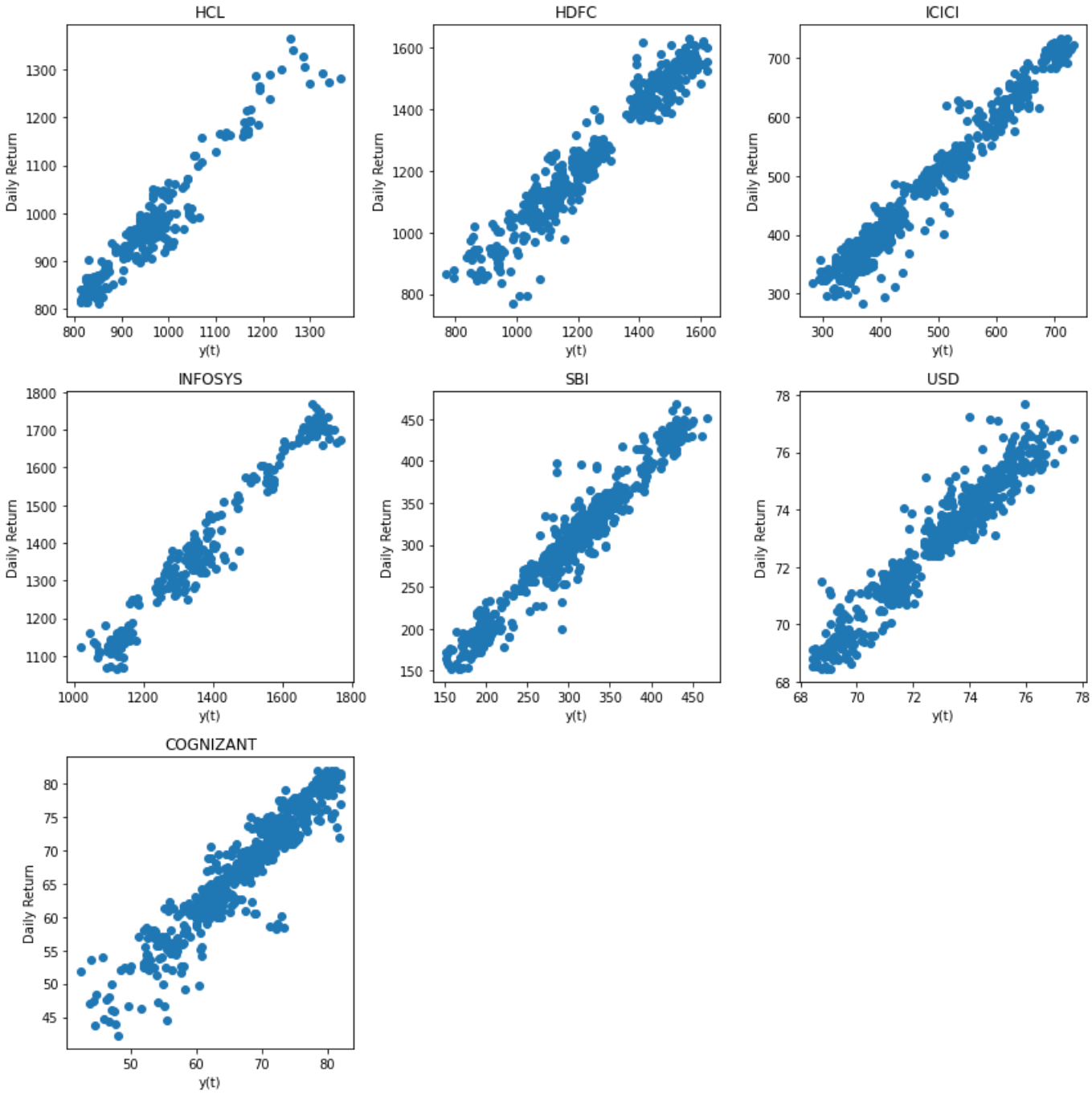

Fig. 14. Lag plot of the 7 stocks with lag = 5

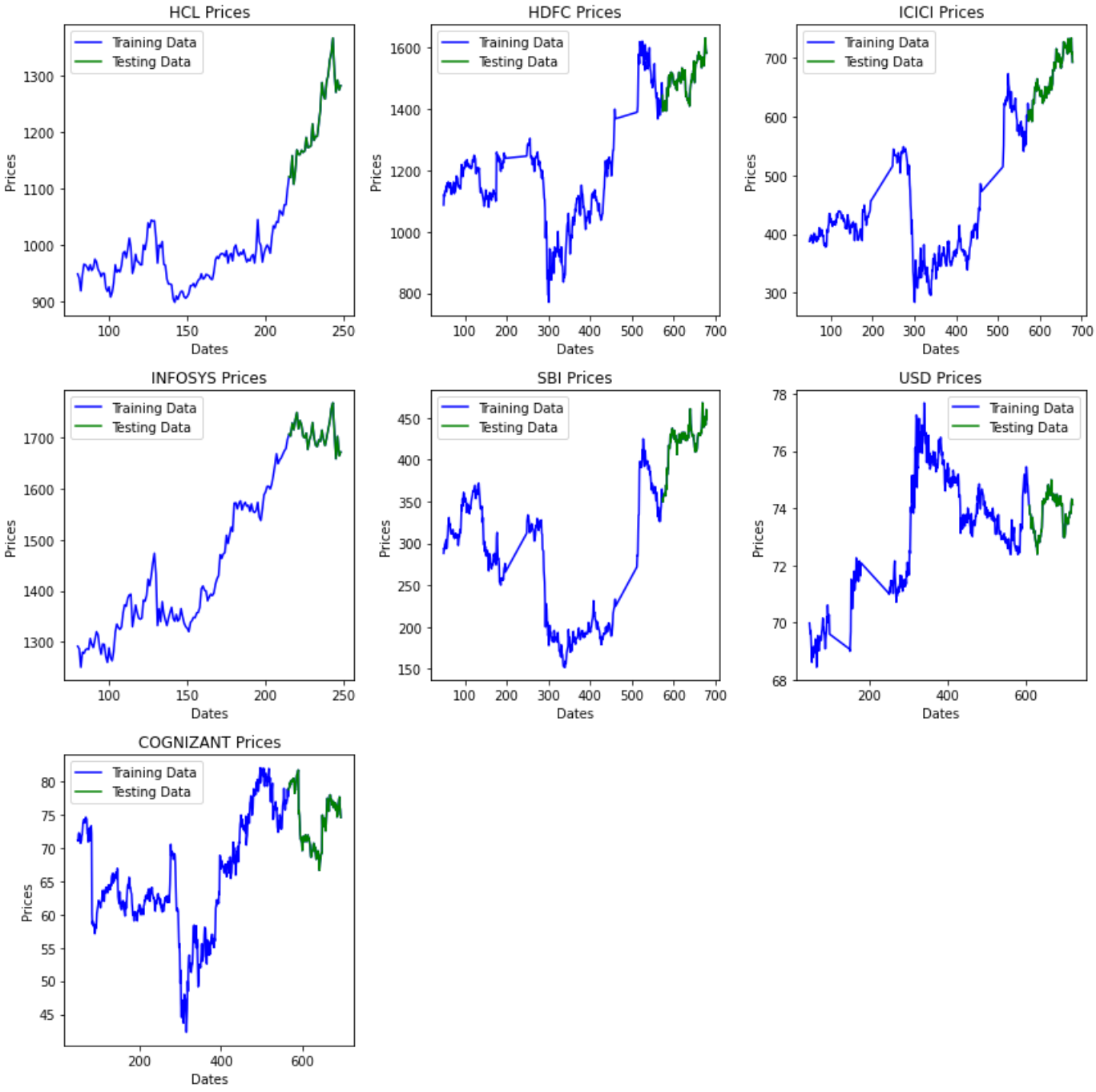

Fig. 15. Training and Validation divisions for each of the 7 stocks

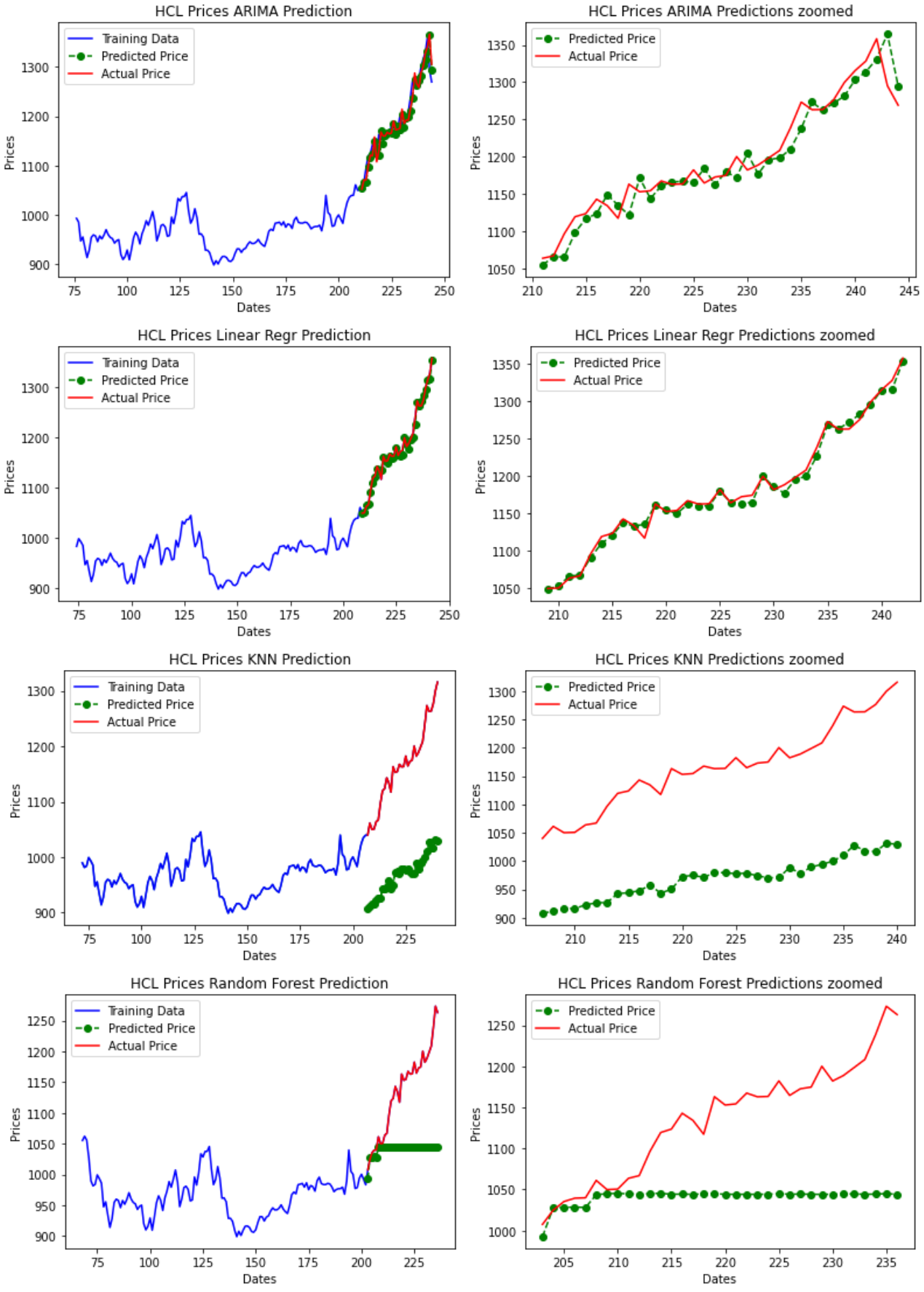

Fig. 16. Predictions for HCL using all methods