Title: A Machine Learning based approach to unleash the impact of COVID-19 on Indian Stock Market

Nusrat Rouf*
Research Lab, Department of Computer Sciences,
Baba Ghulam Shah Badshah University,
Jammu and Kashmir, 185234, India.
Email: nusratrouf@bgsbu.ac.in
**Corresponding author*

Majid Bashir Malik
Department of Computer Sciences
Baba Ghulam Shah Badshah University,
Jammu and Kashmir, 185234, India.
Email: majidbashirmalik@bgsbu.ac.in

Tasleem Arif
Department of Information Technology
Baba Ghulam Shah Badshah University,
Jammu and Kashmir, 185234, India.
Email: t.arif@bgsbu.ac.in

A Machine Learning based approach to unleash the impact of COVID-19 on Indian Stock Market

Abstract

Introduction: Advancement in information technology, be it hardware, software or communication technology, over few decades has rapidly impacted almost every field of study. Machine learning tools and techniques are nowadays applied to every field. It has opened the ways for interdisciplinary research by promising effective analyzation and decision-making strategies. COVID-19 has badly affected more than 200 countries within a short span of time. It has drastically affected both daily activities as well as economic activities. Herd behavior of investors has triggered panic selling. As a result, stock markets around the world have plunged down.

Methods: In this paper, we analyze the impact of COVID-19 on NSE (National Stock Exchange) index Nifty50. We employ Pearson Correlation and investigate the impact of total confirmed cases and daily cases on Nifty50 closing price. We use various machine learning regression models for predictive analysis viz, linear regression with polynomial terms (quadratic, cubic), Decision Tree Regression and Random Forest Regression. Model performance is measured using MSE (Mean Square Error), RMSE (Root Mean Square Error) and R^2 (R Squared) evaluators.

Results: Correlation analysis reveals that total confirmed cases and daily cases in both India and the World have negative correlation with Nifty50 closing prices. Moreover, Nifty50 closing prices are more negatively correlated with total confirmed and daily cases in India. Predictive analysis shows that the Random Forest Regression model outperforms all other models.

Conclusion: We analyze and predict the impact of COVID-19 on closing price of Nifty50 index. We employ Pearson Correlation and investigate the impact of COVID-19 on Nifty50 closing prices. We use various machine learning regression models to predict the closing price of Nifty50 index. Results reveal that the market volatility is directly proportional to increase in number of COVID-19 cases. Random Forest Regression model has comparatively shown better RMSE and R^2 values.

Keywords*: COVID-19, Machine Learning, regression, Nifty50.*

1. Introduction

Novel corona-virus or COVID-19, caused by SARS-CoV-2 initially appeared in Wuhan city of China[1]. On January 5, 2020, WHO (World Health Organization) for the first time published a news regarding a new virus outbreak and within a short span of 25 days, WHO declared the outbreak as a public health emergency of international concern[2]. With its exponential spread, it has affected more than 200 countries and US (United States) has the most confirmed cases[3] .Global death toll has reached to 449 thousand and number of confirmed cases are 10 million as of 28th June, 2020[4]. Till now, there is no specified vaccine for this disease, however several organizations are trying to develop a vaccine including Entos Pharmaceuticals which is developing Fusogenix DNA vaccine, and University of Oxford has developed an adenovirus vaccine, ChAdOx1 nCoV-19. Currently the vaccine is in testing phase[5]. Currently various drugs are used to treat COVID-19 such as *chloroquine*, *hydroxychloroquine* and *Favilavir* [6]. WHO has given some guidelines to prevent spread of this deadly virus that has forced various countries to adopt social distancing, travel restrictions, border shutdowns and many other preventive measures [7][8]. Due to these restrictions, workforce across all sectors has decreased, that in turn has left many people jobless. Lockdown has also tremendously affected the economies of various countries. WHO is continuously inspiring countries across globe to develop smart strategies and methodologies to tackle the ill effects of this pandemic.

The outbreak of this virus has escalated the concerns of all agencies worldwide. Global markets have marked this devastation as *blackswan* event and it has lead them into an uncontrollable and volatile state[9]. It has disrupted and has drastically affected stock market activities worldwide. As soon as the virus started spreading, a sudden decline of prices and increased market volatility was observed. From an economic perspective, it is estimated that in 2020, GDP loss for China could be 6.2% and for United States it could reach to 8.4% [10]. Rest of the world could witness a loss of 5.9%. Withdrawal of money from the markets by the foreign investors is continuous [11]. Panic selling by the investors has made decision making more difficult. The uncertainty in prediction of economy and health is rising due to continuous increase in number of cases and deaths. The figure I below shows the number of confirmed cases for the top five worst affected countries.

This study attempts to measure impact of COVID-19 on stock market outcomes. The paper first summarizes the existing literature on impact of pandemics on stock markets, then explores and analyzes the current available data and attempts to correlate the effect on closing prices of Nifty50 index. It uses different types of regression models to analyze and predict the impact of COVID-19 on Nifty50 index closing price values.

2. Related work

The pandemic has crushed the economy worldwide, although the extent of eventual impact is unknown. While going through the literature we examine that studies have been carried out on relationships between various epidemics and stock markets. [12]examined the impact of SARS on Taiwanese hotel stock prices. [13]investigated the effects of global events on TSE (Toronto Stock Exchange) and KSE (Karachi Stock Exchange). [14]investigated the correlation between H7N9 cases and Chinese stock indices.

[15]studied the impact of COVID-19 on Chinese Stock Market. [16] explored Bayesian regression Model for COVID-19 spread prediction and also studied the impact of COVID-19 of S&P500, [17] analyzed the effect of COVID-19 cases on S&P500 index and [18] studied the effect of COVID-19 on stock market using regression and found a significand relation between COVID-19 variables and stock returns. [19]studied the relationship between pharmaceutical stocks and COVID-19 and reported negative reaction of pharmaceutical stocks. [20]investigated the reaction of stock markets to COVID-19 cases and fatalities. [21]found out that COVID-19 has increased the global financial market risk. [22]reviewed an extensive literature on impact of natural disasters on economy, and also spotlighted that global economic devastation due to COVI19 is unprecedented. [23]analyzed the impact of COVID-19 news on stock market volatility using text analysis and found that COVID-19 made the market more volatile than other similar diseases. The goal of this paper is to analyze and predict the impact of COVID-19 on Nifty50 index closing price values using different types of regression models. The methodology is discussed in next section.

3. Methodology:

The proposed methodology consists of steps from 3.1 to 3.6. In step 3.1, relevant data is collected. In step 3.2, preprocessing of data is done. In step 3.3 feature engineering is employed to extract relevant features. Refined data is analyzed in step 3.4 so as to understand, how sensitive the market is to COVID-19. In step 3.5, correlation analysis is done and step 3.6 discusses various regression models. Visualizations of the results of proposed methodology are mentioned in section 4 of this paper. Figure II shows the overall methodology.

3.1 Data collection:

Stock data as well as corona virus data is openly accessible. Historical daily data for NSE index Nifty50 that measures the weighted average performance of largest 50 companies in NSE was downloaded from www.nseindia.com and global data on COVID-19 was downloaded from https://www.ecdc.europa.eu/en. This data is daily updated for more than 200 corona virus hit countries and regions. Both the datasets were downloaded for a period of five months (30th January 2020 to 22nd June 2020). On 30th January 2020 India received its first corona case. The attribute descriptions for both datasets is mentioned in Table I and Table II.

3.2 Preprocessing:

Data preprocessing is a vital step and purpose is to clean the data for better analyzation. Before data is applied to models, it must be preprocessed so as to get a quality data that can be readily used by machine learning models. Nifty50 dataset was cleaned by removing the instances with missing values that appeared due to the fact that stock market data is not available for weekends and holidays. Data assessment was done to tackle inconsistent stock codes. Duplicate instances were removed from the dataset. Cleaning and noise removal procedures were also applied to COVID-19 dataset. In addition, Feature Aggregations were done so as to make dataset ready for analysis purpose. We also standardized the Date attribute of both datasets and converted the string type to date type. After applying all the filters our COVID-19 dataset contained 25935 observations with 10 attributes and Nifty dataset contained 102 observations and 6 attributes.

3.3 Feature engineering:

From the preprocessed datasets, various features that were critically important, were extracted from both datasets. The selected features were encoded so that data can be readily processed. Feature transformation was done by applying cumulative sum to the Daily_Cases attribute and was named as Total_Cases.

3.4 Data analyzation:

Coronavirus outbreak has affected the stock markets drastically. Increase in the cases worldwide is exponential and has crossed the mark of 10 million. India has entered the list of worst hit countries and number of corona cases are continuously rising.

In this section, we are going to understand how the current crisis impacted stock market. To examine the impact of COVID-19 on Nifty50 index, daily data for India is extracted from COVID-19 dataset. Figure III plots total cases against dates. Figure IV plots closing price of Nifty50 index. It can be inferred from the graph that stock index values have plunged down to 8000 points in the third week of March 2020. We have analyzed the data from 30th January when India received the first COVID-19 positive case.

3.5 Correlation analysis:

We analyze the impact of daily cases and total confirmed cases of corona on Nifty50 closing price. Pearson correlation is used to measure the relationships between variables[24]. The Pearson correlation coefficient r between two variables x and y is calculated as.

$$\mathrm{R}x\,y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Table III shows the correlation matrix for the impact of worldwide daily cases and confirmed cases on Nifty50 closing price. Table IV shows the correlation matrix for the impact of daily cases and confirmed cases in India on Nifty50 closing price. As we can analyze from the below tables closing price (Close) shows the negative correlation with both Daily_Cases and Total_Cases for both world and India using COVID-19 data.

3.6 Regression in Machine learning:

It is one of the most widely used technique of machine learning for predictive modelling. For our methodology we are using quadratic regression, cubic regression, decision tree regression and random forest regression models for predictive analysis.

3.6.1 Linear Regression using polynomial terms:

Linear regression is a supervised machine learning algorithm that models the data linearly. Linear regression with polynomial terms attempts to create a polynomial function that fits the given data[25]. Here the relationship between independent and dependent variable is modeled as a kth degree polynomial.

The generalized formula is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

Here y is independent variable to be predicted and x is independent variable. β_0 is the intercept $\beta_{1 \ldots\ldots} \beta_n$ are set of coefficeints, n is the degree of polynomial and ε is unobserved random error.

The original feature is converted to quadratic and cubic degrees by using the class provided by scikit-learn known as PolynomialFeatures. The data is then trained using linear regression

Linear Regression with Quadratic terms: the generalized model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

For quadratic degree the intercept and coefficients for the observed data are as:

β_0 = 10196.488, β_1 = 0.000000, β_2 = -8.9652, β_3 = 9.1271, ε = 1.74868036e-10

Linear Regression with Quadratic terms: the generalized model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

For cubic degree the observed intercept and coefficients for the observed data are as:

β_0 = 10525.06, β_1 = 0.000000, β_2 = -2.2846, β_3 = 9.1271, ε = -9.38322297e-17

3.6.2 Decision tree regression:

It is a non-parametric supervised machine learning technique and goal is to simplify the complex decision by splitting it[26]. Scikit-learn uses optimized version of CART (Classification and regression trees) algorithm. The algorithm works by forming a tree structure by splitting the features into smaller sub- trees. Impurity criterion is used for splitting of a feature. Scikit-learn uses variance or MAE (mean absolute error) as a measure for impurity.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \mu|$$

Where y is the attribute and n is the number of instances of an attribute and μ is the mean of y_i instances. Predictions are based on what sub-tree a new example will fall into. We applied DecisionTreeRegressor class of scikit-learn for model fitting

3.6.3 Random forest regression:

It is the ensemble machine learning technique where the predictions are made by combining the sequence of decisions made by base learning models[27]. Formally the class of models can be written as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots$$

Where $g(x)$ is the sum of simple base learning models$f_i(x)$. Each base class is a decision tree. We used RandomForest Regressor class of scikit-learn.ensemble module for model fitting.

4. Results and Discussion

The system with 8 GB RAM and 2.6 GHZ processor is used to carry out this work. Python 3.7 with libraries such as sklearn, pandas, numpy, and seaborn are used to perform analysis and prediction. Deeper insights were achieved after carrying out graphical computations. The data split of 80:20 was carried out with 80% data used for training models and 20% data used for testing purpose. Figure V shows the visualized results of applying linear regression with quadratic and cubic terms. Both the models did not properly fit the model. Figure VI shows the visualized results of applying decision tree regression and random forest regression. As it can be inferred from the figure both the models fitted the model in a much better way than the previous quadratic and cubic regression models. The figure VII and VIII show the simplified graph of actual values and predicted values for random forest regression and decision tree regression. Out of the four models random forest regression better fitted the curve.

4.1 Performance Evaluators:

Performance of models was measured using MSE (Mean Square Error), RMSE (Root Mean Square Error) and R^2 (R Square). Table V shows the performance evaluators of all the four models used. It can be inferred from the table that random forest regression has lower MSE, RMSE and higher R^2

5. Conclusion:

COVID-19 is generating shock waves around the world by affecting almost every sector. It has crushed the economy worldwide. We analyze and predict the impact of COVID-19 on closing price of Nifty50 index. We use machine learning regression algorithms to predict the closing price of Nifty50 index and use various performance measures to evaluate the performance of models. The results reveal that Nifty50 index is sensitive to corona virus disease. The market volatility is directly proportional to increased COVID-19 cases. Random Forest Regressor has comparatively shown better RMSE and R^2 values. This work can be extended by considering various other variables such as deaths, recovered cases and analyzing the impact on stock markets. Further effect of COVID-19 on stock market returns can be considered.

Compliance with the Ethical standards: No funding was received for this work.

Ethical Approval: this article does not contain any studies with human participants or animals performed by any other authors.

6. References:

[1] F. Wu *et al.*, “A new coronavirus associated with human respiratory disease in China,” *Nature*, vol. 579, no. 7798, pp. 265–269, 2020.

[2] “Coronavirus disease 2019.” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. [Accessed: 30-Jun-2020].

[3] “COVID-19 Map - Johns Hopkins Coronavirus Resource Center.” [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 30-Jun-2020].

[4] “Coronavirus Update (Live): 10,436,954 Cases and 508,876 Deaths from COVID-19 Virus Pandemic - Worldometer.” [Online]. Available: https://www.worldometers.info/coronavirus/?zarsrc=130. [Accessed: 30-Jun-2020].

[5] “Coronavirus outbreak: Top coronavirus drugs and vaccines in development.” [Online]. Available: https://www.clinicaltrialsarena.com/analysis/coronavirus-mers-cov-drugs/. [Accessed: 09-Jul-2020].

[6] “35 drugs in the race to treat new coronavirus, Health News, ET HealthWorld.” [Online]. Available: https://health.economictimes.indiatimes.com/news/pharma/35-drugs-in-the-race-to-treat-new-coronavirus/74525606. [Accessed: 14-Jul-2020].

[7] “Centers for Disease Control and Prevention.” [Online]. Available: https://www.cdc.gov/. [Accessed: 30-Jun-2020].

[8] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, “Machine learning based approaches for detecting COVID-19 using clinical text data,” *Int. J. Inf. Technol.*, pp. 1–9, 2020.

[9] L. Morales and B. A. Callaghan, “Covid-19 -Global Stock Markets " Black Swan " Critical Letters in Economics & Finance Covid19 : Global Stock Markets ‘ Black Swan ,’” no. March, 2020.

[10] W. McKibbin and R. Fernando, “The Global Macroeconomic Impacts of COVID-19,” *Brookings Inst.*, no. March, pp. 1–43, 2020.

[11] A. Khanthavit, “Foreign Investors ’ Abnormal Trading Behavior in the Time of COVID-19,” no. June, 2020.

[12] M. H. Chen, S. C. (Shawn) Jang, and W. G. Kim, “The impact of the SARS outbreak on Taiwanese hotel stock performance: An event-study approach,” *Int. J. Hosp. Manag.*, vol. 26, no. 1, pp. 200–212, 2007.

[13] W. H. Yeung and A. Aman, “Sensitivity of stock indices to global events: the perspective for Pakistani Canadians,” *J. Econ. Adm. Sci.*, vol. 32, no. 2, pp. 102–119, 2016.

[14] W. Sun, “H7N9 not only endanger human health but also hit stock marketing,” *Adv. Dis. Control Prev.*, vol. 2, no. 1, p. 1, 2017.

[15] A. M. Al-awadhi, K. Alsaifi, A. Al-awadhi, and S. Alhammadi, “Death and contagious infectious diseases: Impact ofthe COVID-19 virus on stock market returns ,” no. January, 2020.

[16] B. M. Pavlyshenko, “Regression Approach for Modeling COVID-19 Spread and its Impact On Stock Market,” vol. 2020, no. 1, pp. 1–10, 2020.

[17] H. Yilmazkuday, “COVID-19 Effects on the S & P 500 Index,” vol. 2019, no. March, pp. 1–15, 2020.

[18] F. ZEREN and A. HIZARCI, “the Impact of Covid-19 Coronavirus on Stock Markets: Evidence From Selected Countries,” *Muhasebe ve Finans İncelemeleri Derg.*, vol. 1, pp. 78–84, 2020.

[19] M. Aravind and C. G. Manojkrishnan, “COVID 19: Effect on leading pharmaceutical stocks listed with NSE,” *Int. J. Res. Pharm. Sci.*, vol. 11, no. Special Issue 1, pp. 31–36, 2020.

[20] B. N. Ashraf, “Stock markets reaction to COVID-19: cases or fatalities?,” *Res. Int. Bus. Financ.*, no. May, p. 101249, 2020.

[21] D. Zhang, M. Hu, and Q. Ji, “Financial markets under the global pandemic of COVID-19 ,” no. January, 2020.

[22] J. W. Goodell, “COVID-19 and finance: Agendas for future research ,” no. January, 2020.

[23] S. R. Baker, N. Bloom, J. Davis, K. Kost, M. Sammon, and T. Viratyosin, “The unprecedented stock market reaction to Covid-19,” *PANDEMICS LONG-RUN Eff. Eff.*, vol. 1, no. DP 14543, pp. 33–42, 2020.

[24] W. Kirch, Ed., “Pearson’s Correlation Coefficient BT - Encyclopedia of Public Health,” Dordrecht: Springer Netherlands, 2008, pp. 1090–1091.

[25] B. Sun, H. Liu, S. Zhou, and W. Li, “Evaluating the Performance of Polynomial Regression Method with Different Parameters during Color Characterization,” *Math. Probl. Eng.*, vol. 2014, p. 418651, 2014.

[26] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, “Decision tree regression for soft classification of remote sensing data,” *Remote Sens. Environ.*, vol. 97, no. 3, pp. 322–336, 2005.

[27] Y. L. Pavlov, “Random forests,” *Random For.*, pp. 1–122, 2019.

Tables:

Table I: The attribute/column description of Nifty50 data set

Column Name	Description
Date	Date of a particular day
Open	Opening price of a particular stock
High	Highest price of a particular stock in a day
Low	Lowest price of a particular stock in a day
Close	Closing price of a particular stock at the end of a day
Volume	Number of stocks traded in a day for a particular stock

Table II: The attribute/column description for global COVID-19 dataset

Column Name	Description
Date	Date of a particular day
Daily_Cases	Total number of cases in a day
Deaths	Total number of deaths in a day
Country	Country
GeoId	Geographic Identification of a country
CountryAndterritoriesCode	Code of a particular country
PopData2019	Population of a country as per 2019 census

Table III: Correlation Matrix (World)

	Close	Daily_Cases	Total_Cases
Close	1.0	-0.21294	-0.14343
Daily_Cases	-0.21294	1.0	0.98231
Total_Cases	-0.14343	0.98231	1.0

Table IV: Table IV: Correlation Matrix (India)

	Close	Daily_Cases	Total_Cases
Close	1.0	-0.58852	-0.34186
Daily_Cases	-0.58852	1.0	0.92228

Total_Cases	-0.34186	0.92228	1.0

Table V: Performance Evaluation of regression models

Performance Evaluators	Quadratic Regression	Cubic Regression	Decision Tree Regression	Random Forest Regression
MSE	704482.93	669565.95	30742.455	15282.644
RMSE	839.33	818.27	175.335	123.62
R^2	0.24	0.27	0.951	0.973

Figures:

Figure I: Top listed five countries with highest number of total confirmed cases

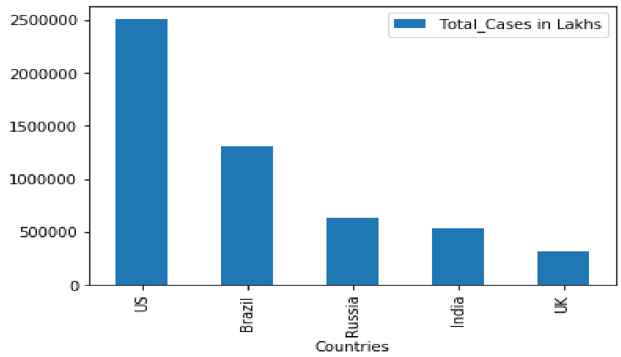


Figure II: Methodology

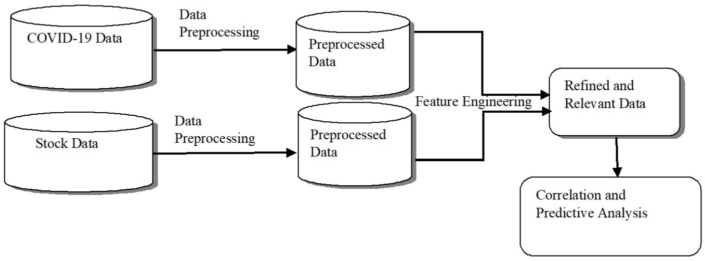


Figure III Total number of cases in India and the World

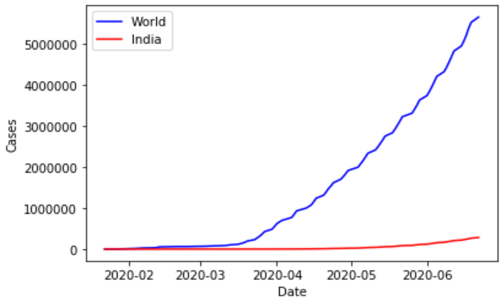


Figure IV: Nifty50 Closing Price

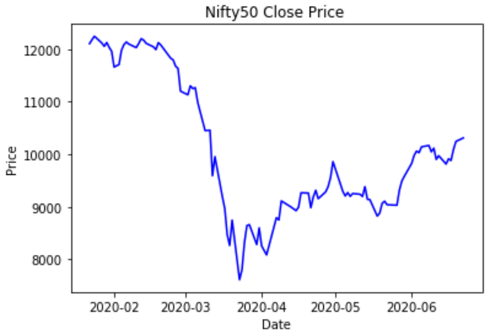


Figure V: Quadratic and cubic Regression actual vs predicted values

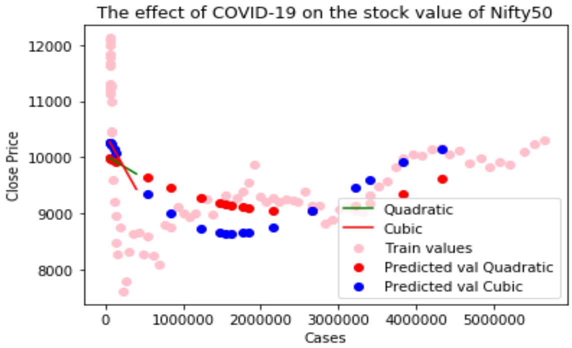


Figure IIII: Random Forest and Decision Tree Regression actual vs predicted

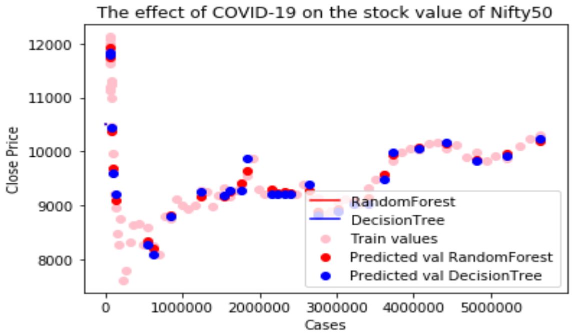


Figure IVI: Decision Tree Regression actual vs predicted values

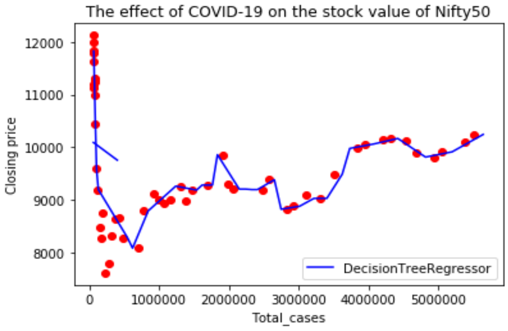
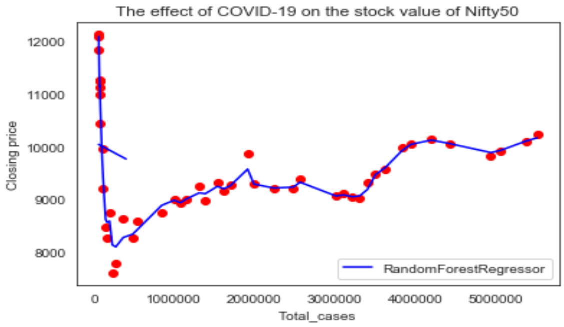


Figure VI: Random Forest Regression actual vs predicted values


The effect of COVID-19 on the stock value of Nifty50
Closing price
Total_cases
RandomForestRegressor