

A COMPARISON OF STOCK PRICE PREDICTION TECHNIQUES: OLD SCHOOL VERSUS NEW SCHOOL

Abstract : Stock market forecasting is an activity that has fascinated analysts and investors alike for decades. The uncertainty associated with investor behavior or rather, human behavior in general, that is owed to a multitude of physical, psychological, and financial facets of decision making, makes forecasting an exasperating and challenging task to undertake. Theory postulates the notion that market efficiency is inherent and individual attempts to disprove this hypothesis are futile. To check the feasibility of the claims being put forth by a growing number of analysts and investors who claim to contradict the 'Efficient Market Hypothesis', we compare different techniques of stock price forecasting. In the paper, we construct our own OLS, SVR, and LSTM neural network models through statistical software to test the best among them.

Analysis for different time frames i.e. 20 years, 6 years and 1 year using NSE's historical index yielded the result that OLS is the most efficient model in the longest time frame (20 years), followed by LSTM and SVR respectively. We also observed that SVR performs extremely poorly across all time frames. We extend the scope of this study to four companies listed on NSE to check the viability of our models. It was observed that company based analysis yielded more or less the same result as Index based analysis.

I. INTRODUCTION :

The mechanism of financial markets ensures the channelisation of funds from the savers to the investors, thereby becoming an enabler of economic growth. Pertaining to this attribute of the markets, theorists and researchers have been trying to understand how the market operates for a very long time. However, the task is not a child's play because of the chaotic nature of the markets. Various factors, including the political and economic environment of the country, behavioral traits and other butterfly effects collectively influence the market and it becomes increasingly difficult to assess the impact of any of these factors individually.

The intensity of stock market efficiency has been an extensively talked-about topic since the early 1960's. Stepping on the shoulder of giants, *Fama(1970)* performed a theoretical and empirical study on the efficiency of markets as propagated by earlier theorists *Bachelier (1900)*, *Mandelbrot (1963)*, and *Samuelson (1965)* in the form of Random Walk Hypothesis. The theory straightforwardly claims that market prices fully reflect the information and thus, all attempts to try and beat the market for earning higher returns are futile. Financial researchers, analysts, and investors worldwide have now and then refuted this theory by asserting that the market is less noisy and more smooth in the long run, making the task of forecasting achievable. Leigh et al. (2008) defied the Efficient Market

Hypothesis by constructing a pattern-recognizing algorithm using **Bull flag pattern** and successfully proved the superiority of the model in beating average market profit for 9000 trading days of NYSE close price. Several such studies have been undertaken for the past two decades and the process still continues with the introduction of Artificial Intelligence in the field of finance.

Through this paper, we aim to construct three variants of stock prediction models and compare their efficiency in forecasting Closing prices of company stocks and index.


In section **II** of this paper, we provide a background on the existing literature and studies in the field of Quantitative trading. Section **III** describes in detail the data used to undertake our study and the methodologies followed for model-development. In section **IV**, we present our empirical findings and attempt to find the reasons thereof. With section **V**, we conclude our discussion and give some policy recommendations. Section **VI** consists of references.

II. LITERATURE REVIEW :

Academia is flooded nowadays by the vast stream of researches and developments in the arena of stock prediction. While some of them are based on statistical methods and techniques, others stem from the use of Artificial Intelligence.

Bhuriya et al. (2017) tried to test the validity of various types of regression techniques to forecast stock prices of Tata Consultancy Services stocks using variables like Open, High, Low, Close and Volume. Out of the three models tested, namely Linear, Polynomial and Radial Basis Function, the Linear model turned out to be the best with the confidence value as high as 0.97.

Roondiwala et al. (2017) executed the LSTM neural network model to forecast Nifty 50 closing prices using the Open-High-Low-Close chart. The results were exceptional in the sense that their model got an RMSE of only 0.00859 for daily close price test data.

Zhang et al. (2018) used Decision Tree C.5 algorithm to formulate a classification model for forecasting stock returns and classifying stocks according to their returns for the Chinese stock market. **Their model yielded accuracy as high as 98% for better performing stocks and low accuracy (almost 0%) for poorly performing stocks,** indicating an asymmetry of model performance in anticipating risk (downside return) and return (upside return). 

Lijuan Cao et al. (2001) performed a comparative study of SVM and Neural network based on **Back-propagation** using S&P 500 Daily price index. The study concludes that SVM yielded superior results over the Neural network model in predicting the stocks.

The objective of most of these researches and studies was to develop models which can reshape the methods of forecasting. Meanwhile, this paper aims at developing three models for stock price prediction:

- Traditionally used models - OLS (Ordinary Least Squares) Linear regression model and SVR (Support Vector Regression) model
- Modern Artificial Intelligence model - LSTM (Long Short Term Memory) network model in Deep learning

III. DATA AND METHODOLOGY :

“In God we trust, all others must bring data.” - William Edwards Deming

1. **Data** - For our study, we collected historical price data on Daily basis on features like *Open, Close, High, Low* and *Volume* from NSE Equity Stock Index *NIFTY50* from 1st January, 2000 to 31st March, 2020. This Index was hand-picked by keeping in mind its exceptional characterization of the Indian financial markets. Moreover, we selected NSE data for our analysis because it ranks 11th (as per April 2018) among the world's largest stock exchanges in terms of Market Capitalization. For company based analysis, we obtained data on four randomly selected companies (representing industries like, Banking, Telecom, Consultancy and Pharmaceuticals) listed on NSE, namely '*HDFC Bank*', '*Reliance Industries Limited*', '*Tata Consultancy Services Limited*' and '*Sun Pharmaceuticals Industries Limited*' from the Yahoo financial website.

To perform training and testing of our models, we have divided our time series data into three varying sets of time frames: 20 years (1st January, 2000 to 31st March, 2020), 6 years (1st January 2014 to 31st March, 2020) and 1 year (1st January, 2019 to 31st March, 2020). This dissection of data is being done to assess the efficiency of our models in different time spans and to examine whether the models lose their accuracy as smaller and smaller time periods are taken.

2. **Construction of the models** - We used R version 3.6.3 and Python coding to develop our statistical and machine learning models to predict Close Prices of NIFTY50 Index and the Equity stocks of the companies selected. In all three of them, we practiced *Feature Engineering* i.e. train the model based on certain features (variables), test the predictions and re-run the model on another set of features to minimize the error and eventually keep the features in the model that accord the least error. Moreover, we split the data into two parts - 70% for training the model and rest 30% for testing our predictions.

- ❑ **OLS** - We employed Feature Engineering through Backward Selection technique i.e. beginning with all the features in data and then continue dropping irrelevant variables till the best model is formed. We used Akaike Information Criterion (AIC), Random Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Variance Inflation Factor (VIF) and Adjusted R-Square to compare among the different versions of regression models.

While running the codes, we identified some peculiarities in our model. The Normal QQ-plot (which shows whether the data is normally distributed or not) indicated the presence of outliers for the Long time frame. However, this attribute got eliminated on its own as the time frame became smaller and smaller and the Normal QQ-plot became approximately linear.

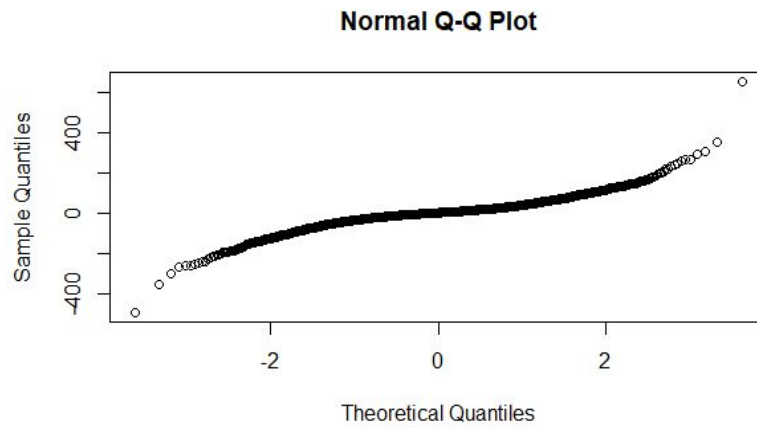


Fig. 1 : Normal QQ-plot for NIFTY50 2000-2020

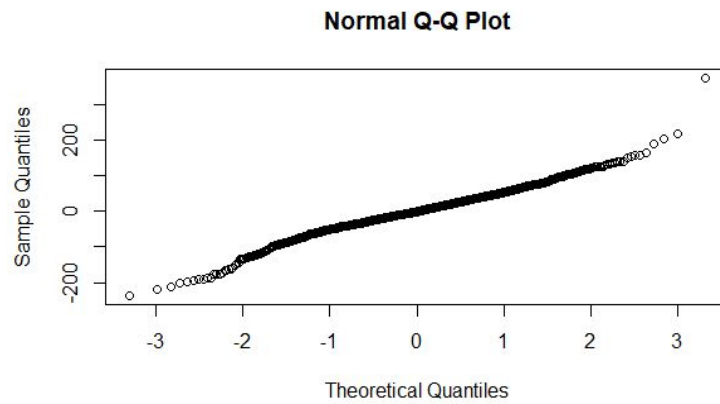


Fig. 2 : Normal QQ-plot for NIFTY50 2014-2020

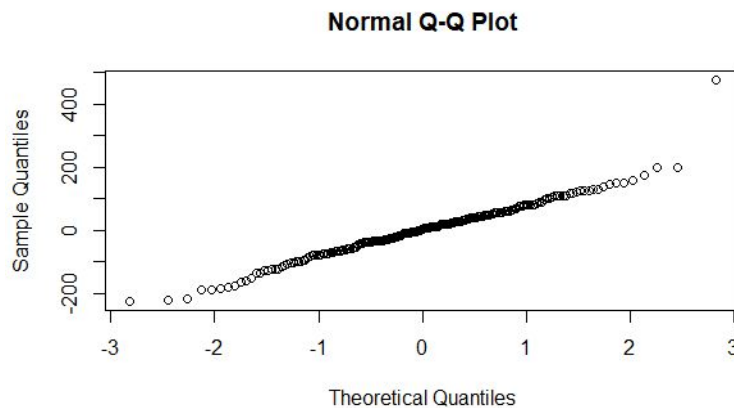


Fig. 3 : Normal QQ-plot for NIFTY50 2019-2020



The reason behind this anomaly appears to be the time-series nature of our data. Our final model was based on two features, namely 'Open' and 'Volume'.

- ❑ **SVR** - Our SVR model has been fabricated via the Tuning process, which is based on kernel functions (linear, polynomial, sigmoid or radial-basis). The process includes training the various models with varying tolerable error (epsilon) and associated cost parameters and running the simulation multiple times using sampling to choose the combination of error and cost parameters for the kernel function. We have used Radial-basis kernel function, an automatically generated kernel function made available by R 3.6.3 which is defined as:

$$K(x, x') = \exp (-\| x - x' \|^2 / 2\sigma^2)$$

where, x and x' are two feature vectors and σ is a free hyperparameter that the tuning process tries to estimate. The acumen of this model of machine learning lies in the fact that it is capable of assigning penalty, on using an additional feature, using a cost parameter to avoid over-fitting.

- ❑ **LSTM Network** - First of all, we transformed the time-series data by normalizing it in order to fasten the learning speed of our neural network using the following formula for Min-Max normalization:

$$z = \{ x - \min(x) \} / \{ \max(x) - \min(x) \}$$

LSTM is a sequential model which uses Back-propagation i.e. assign posterior weights to variables, given the knowledge of error that arises due to prior weights, and repeat these iterations several times in order to tune the final weights in such a way that error gets minimized. Our model uses Mean squared error (MSE) for the error attribute.

We have used Keras library from Python to train our model, which makes use of Sigmoid as its activation function.

$$\sigma(x) = 1 / (1 + e^{-x}) \quad : \text{ Sigmoid function}$$

IV. EMPIRICAL FINDINGS :

After running the time-series data of NIFTY50 and the Four randomly selected companies listed on NSE, we obtained the following results:

- ❑ **NIFTY50 :-**

The Adjusted R^2 for the data emerged to be higher than 0.9 for all of the time frames taken into consideration, indicating towards a very good fit.

Table 1 gives the Adjusted R^2 values for NIFTY50 and all of the four companies.

Table 1 : Adjusted R²

Time Frame	NIFTY50	RELIANCE	TCS	SUN PHARMA	HDFC BANK
2000-2020	0.9991	0.9985	0.9993	0.9994	0.9993
2014-2020	0.9971	0.9976	0.9783	0.9924	0.9990
2019-2020	0.9525	0.9537	0.9326	0.8815	0.9647

This finding is further supported by the plotting of actual and predicted values of the closing prices of NIFTY50 in figure .

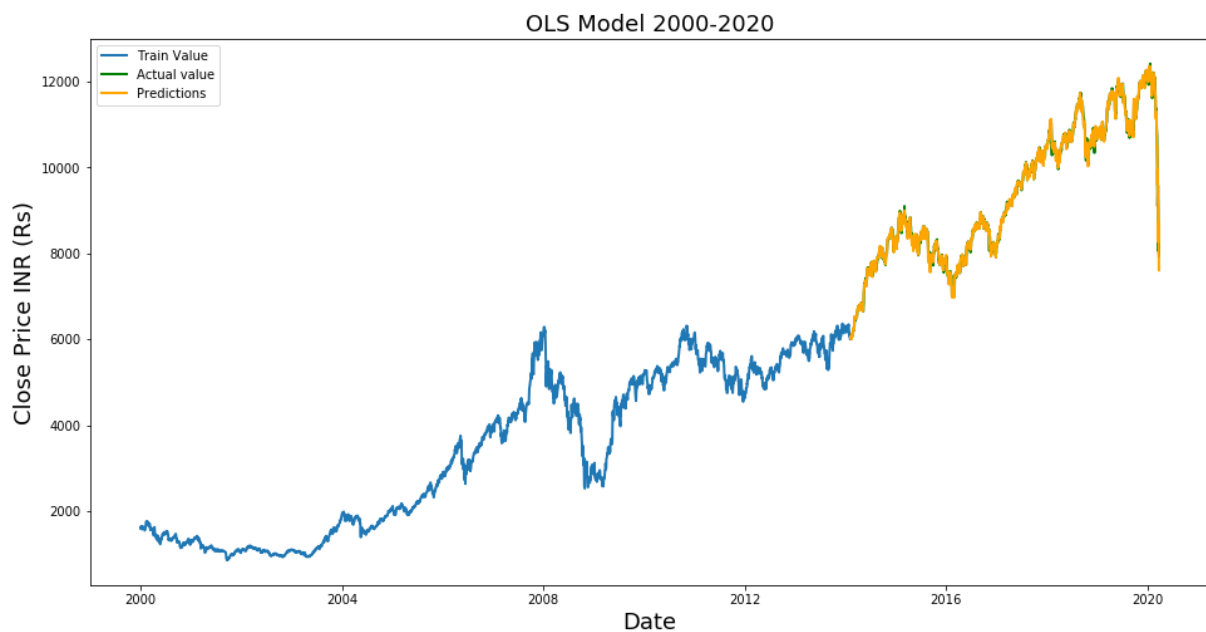


Fig. 4 : NIFTY50 OLS Prediction for 2000-2020

However, we found that the OLS forecasting becomes weaker as the time frame becomes smaller, which is evident from the following graphs (Fig. 5 and 6):

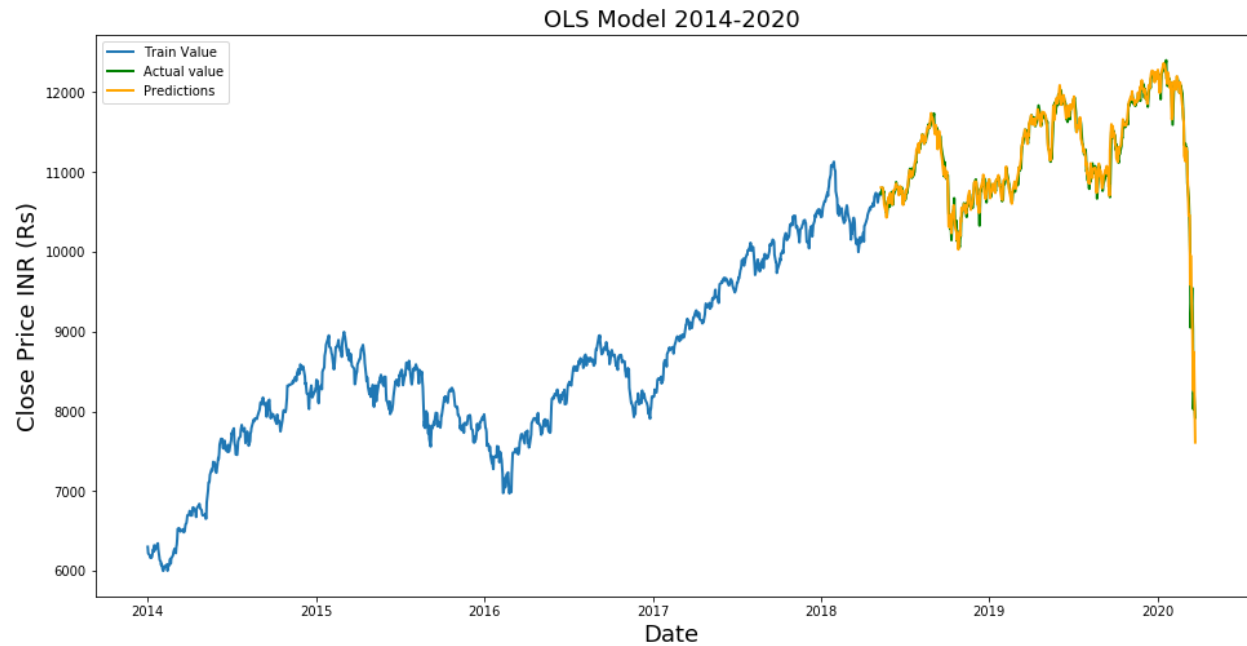


Fig. 5 : NIFTY50 OLS Prediction for 2014-2020

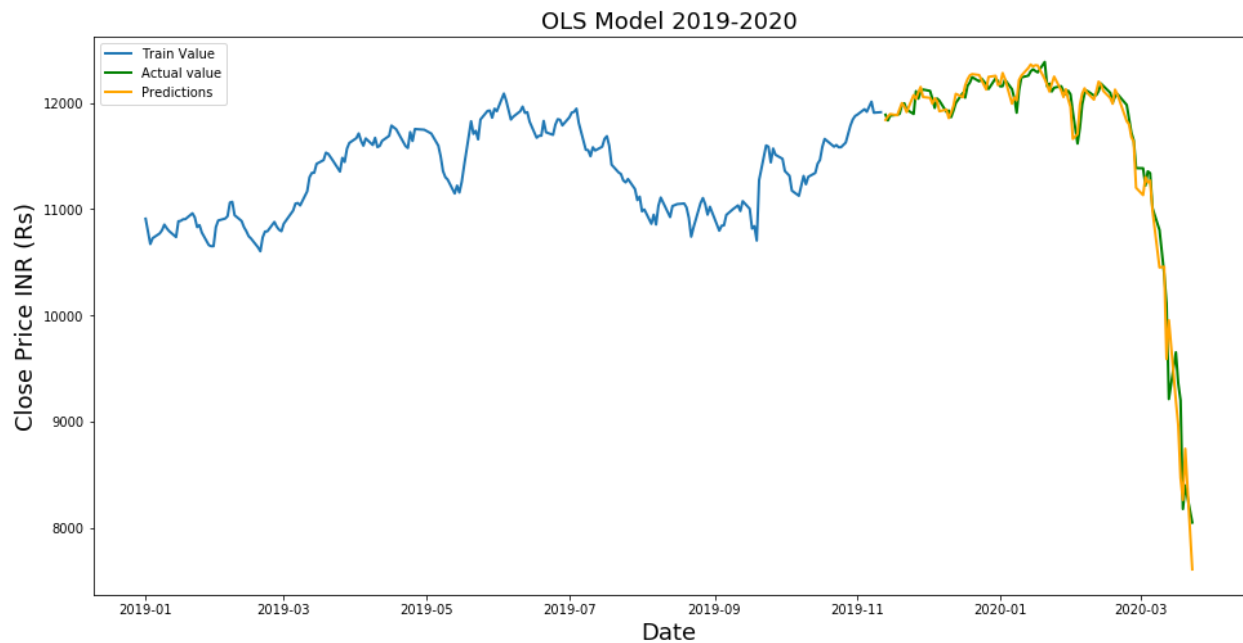


Fig. 6 : NIFTY50 OLS Prediction for 2019-2020

The LSTM Network model yields the same pattern as the OLS model when tested for different time frames (Fig. 7-9). Predicted values are closer to actual values for larger time frames and the model almost completely loses out its accuracy for one year long time span, as can be seen in the figures below.



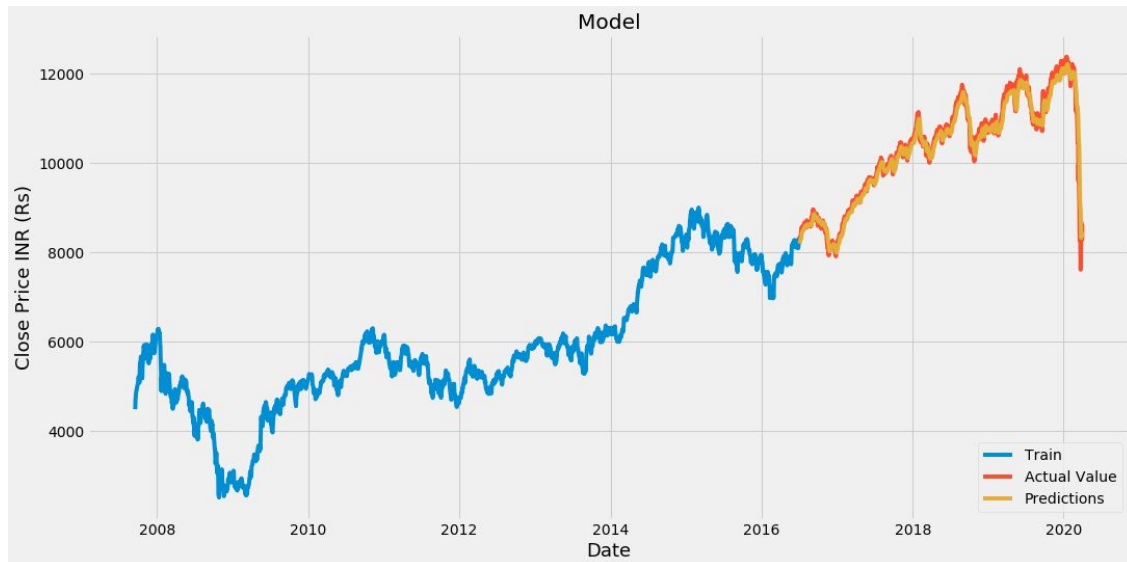
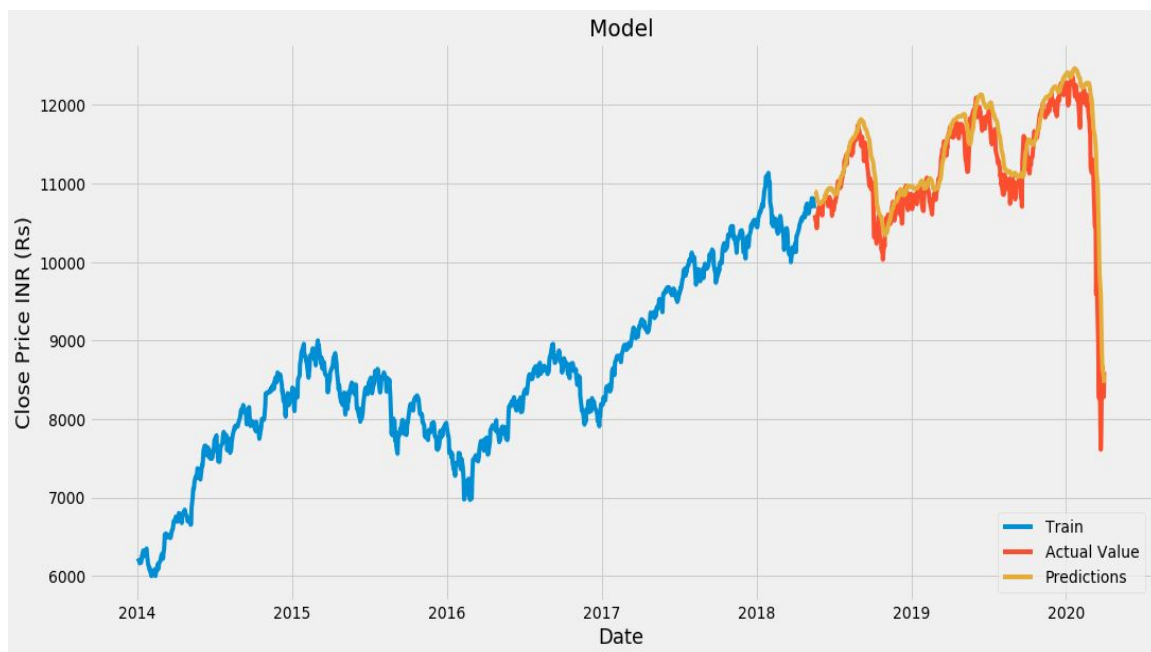


Fig. 7 : NIFTY50 LSTM Prediction for 2000-2020

Fig. 8 : NIFTY50 LSTM Prediction for 2014-2020



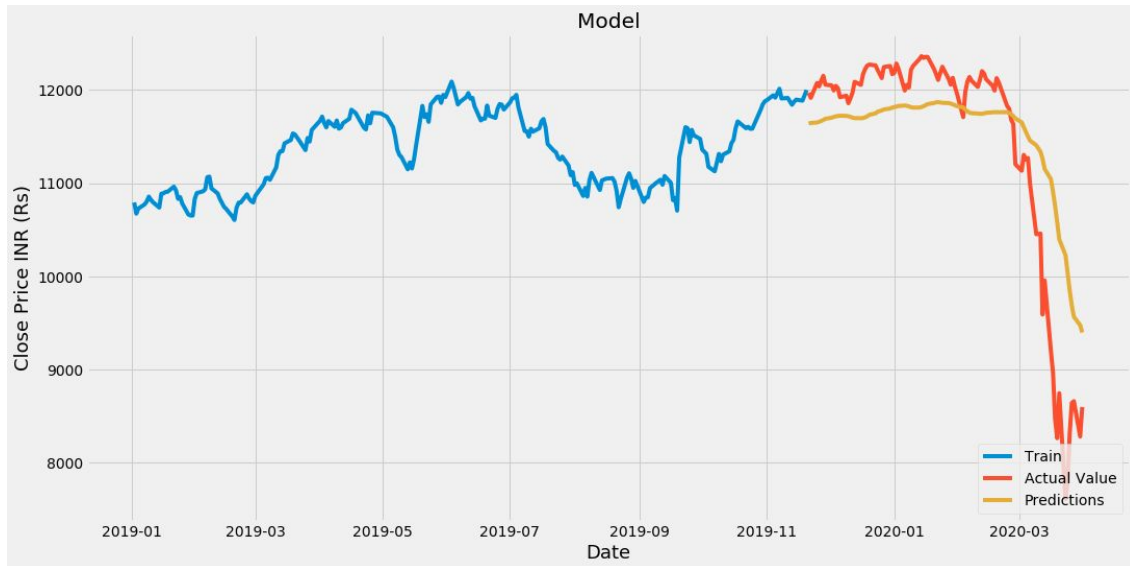


Fig.9 : NIFTY50 LSTM Prediction for 2019-2020



This finding is not surprising since LSTM Network models operate better on large data sets because later neurons possess lesser memory than early ones. Thus, the results are not distortionary and hence, our model is not reflecting any anomaly.

The SVR model shows extremely poor results, both in absolute as well as relative terms. The forecasted values form a mirror image of the actual values and the entire prediction loses the trend of the actual variable.

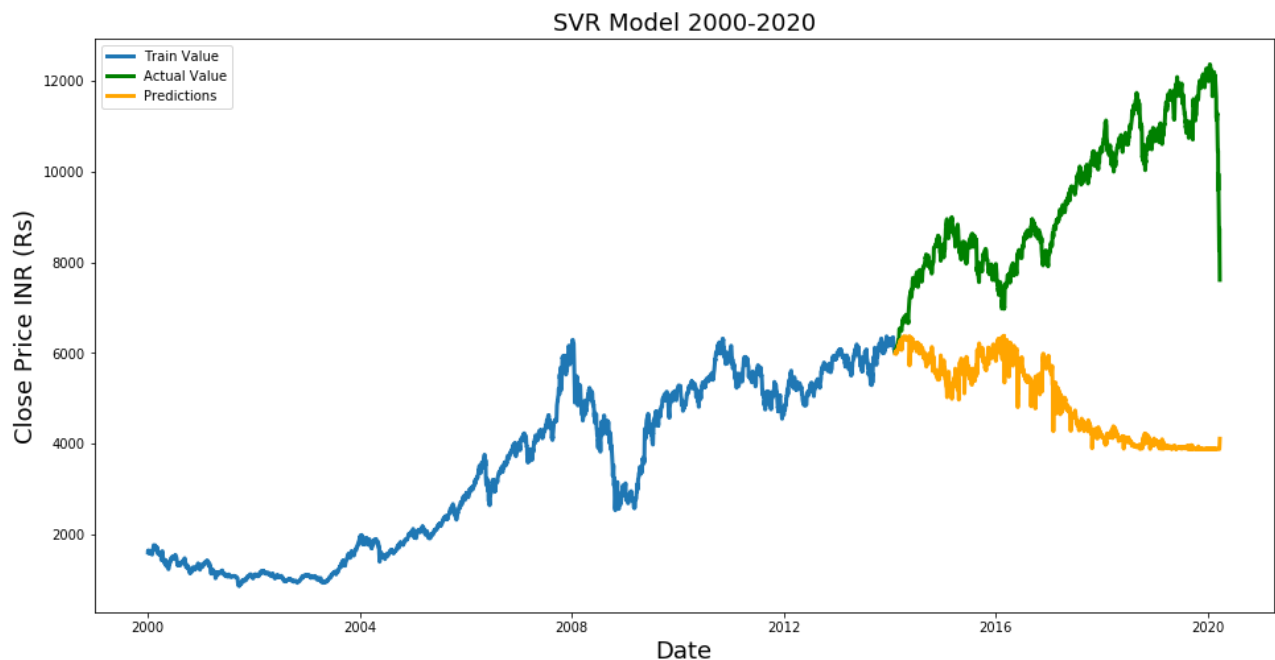


Fig. 10 : NIFTY50 SVR prediction for 2000-2020

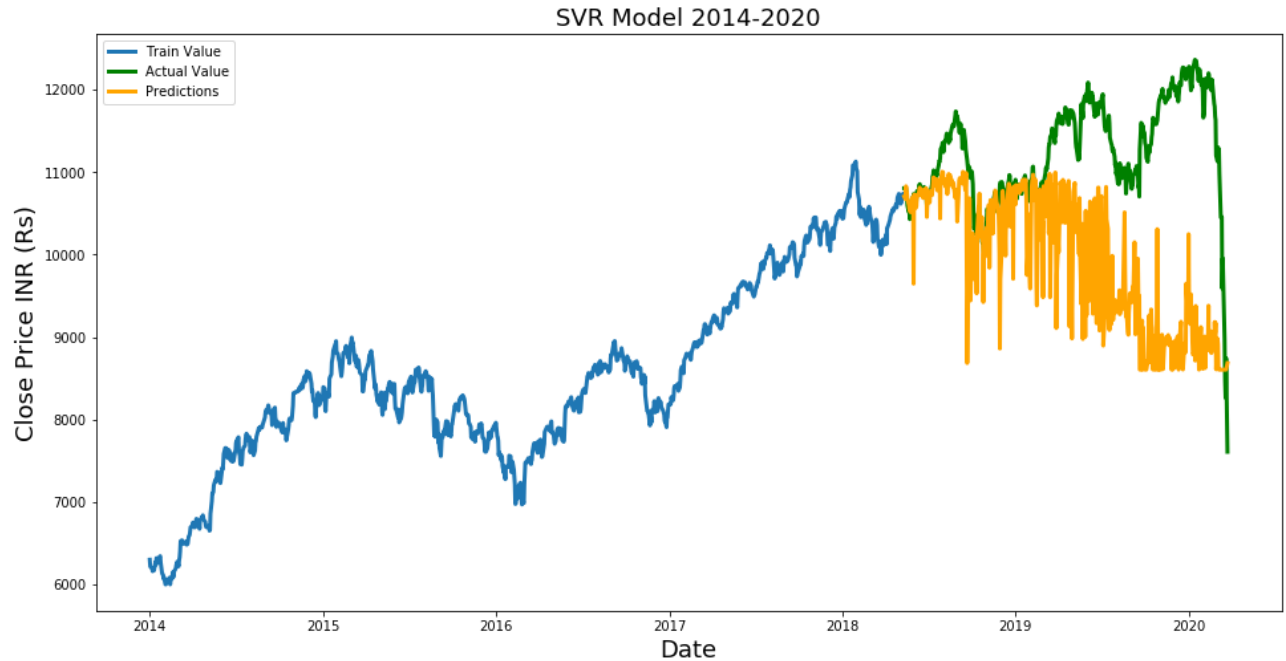


Fig. 11 : NIFTY50 SVR prediction for 2014-2020

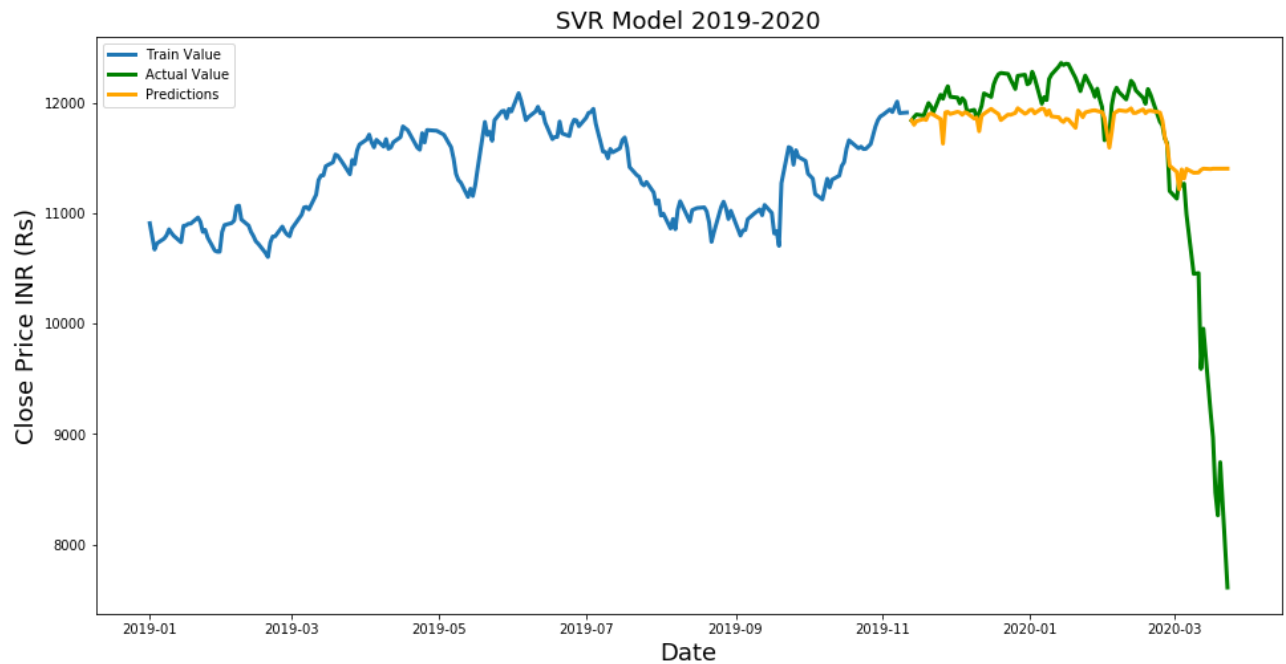


Fig. 12 : NIFTY50 SVR prediction for 2019-2020

For comparing the three models, we collate the Root mean squared error (RMSE) and Mean absolute percentage error (MAPE) for each and compare it across the different time periods as well. The following tables provides the findings:

Table 2 : NIFTY50 RMSE

Time Frame	OLS	SVR	LSTM
2000-2020	77.96	5054.32	194.02
2014-2020	106.93	1763.75	281.69
2019-2020	173.41	824.47	813.57

Table 3 : NIFTY50 MAPE

Time Frame	OLS	SVR
2000-2020	0.0057	0.4413
2014-2020	0.0066	0.1142
2019-2020	0.0099	0.042

As we can see, OLS outperforms both LSTM and SVR in terms of error minimization across all time spans. However, the errors for OLS get bigger as the time frame becomes smaller, yielding the same result as the graphs and adjusted R^2 . Errors for LSTM follow the same pattern. SVR yields extremely high errors, which is also being reflected in the graphs as the model fits the data abysmally. The striking feature, however, is the downward trend in errors for SVR. As time frames contract, SVR becomes less poor in absolute terms. Relatively, SVR is still a poorer model as compared to OLS and LSTM both. The reason behind this could be the possibility of linear separation of features in the data while building SVR which defeats the very purpose of using the model as an improvement over Simple linear regression. Thus, there exists the problem of over-fitting in our SVR model.

❑ Companies :-

Data on companies more or less offered the same results as NIFTY50 data. The major trend of all of the prediction models losing their efficiency as time span contracts is being iterated by this data as well. Comparison of errors (RMSE and MAPE) across the three time periods for the three models gives out the same result.

- 1) **Reliance** - Historical data on Reliance Industries ltd. showed the exact same trends as in the NIFTY50 in terms of every aspect of analysis that we have undertaken so far for the latter.

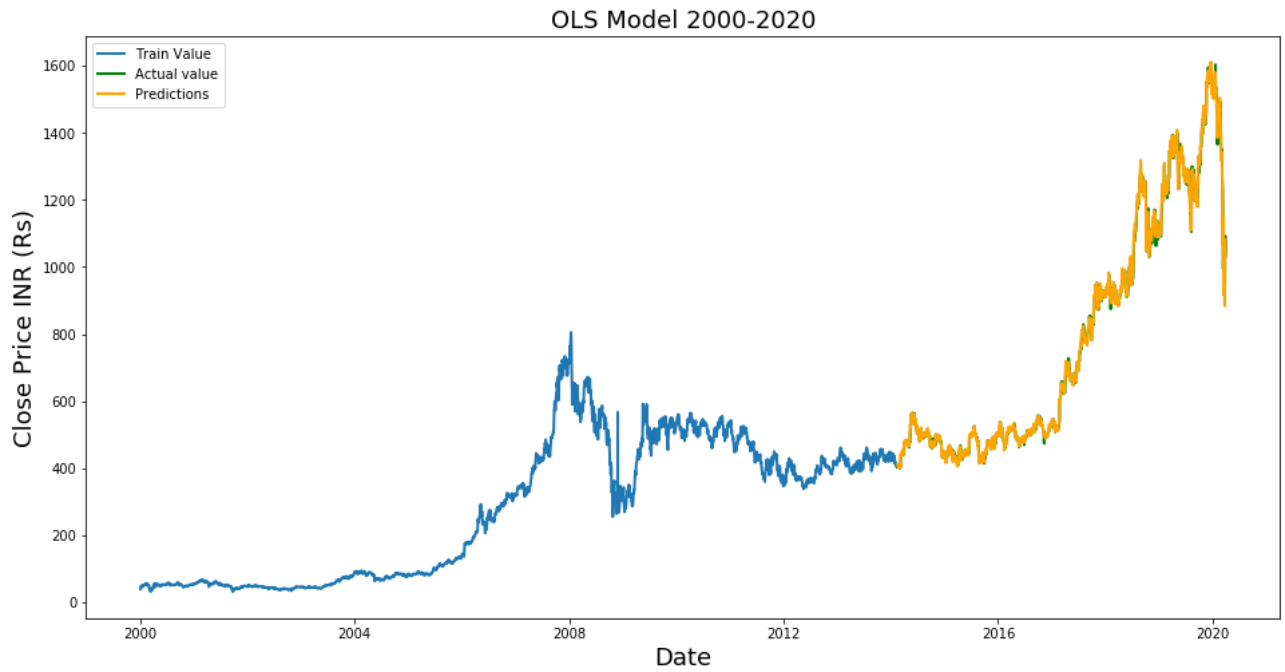


Fig. 13 : Reliance OLS Prediction for 2000-2020

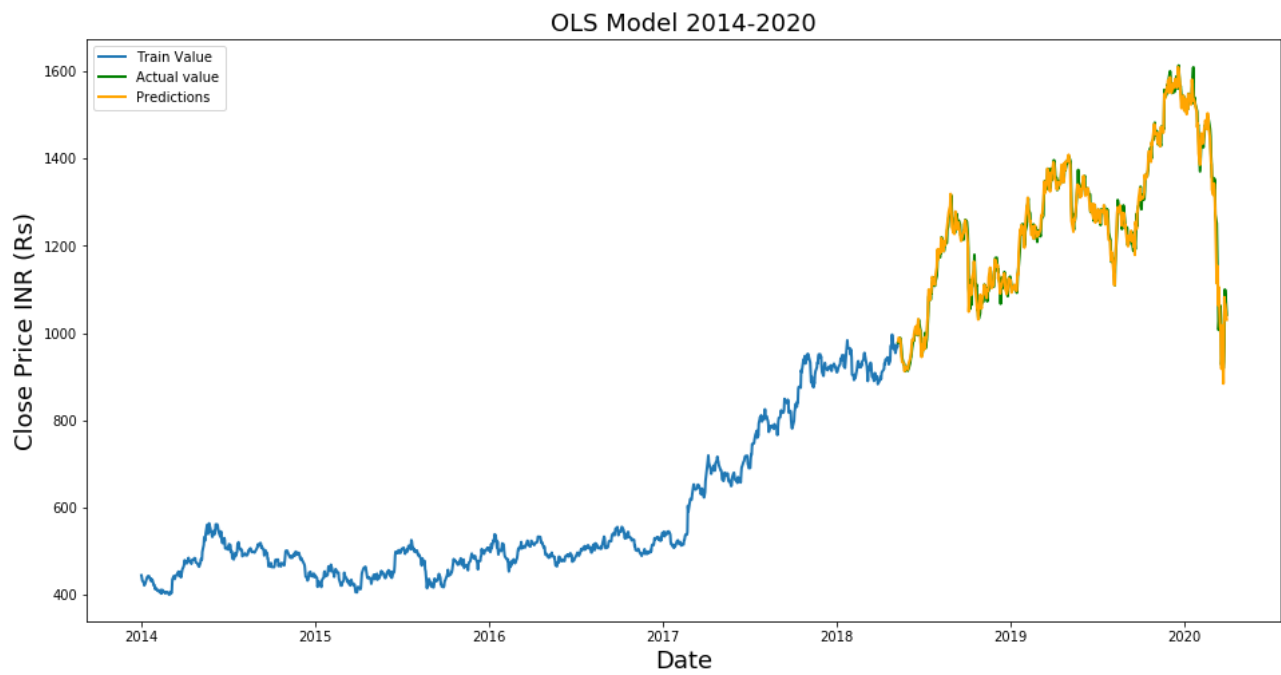


Fig. 14 : Reliance OLS Prediction for 2014-2020

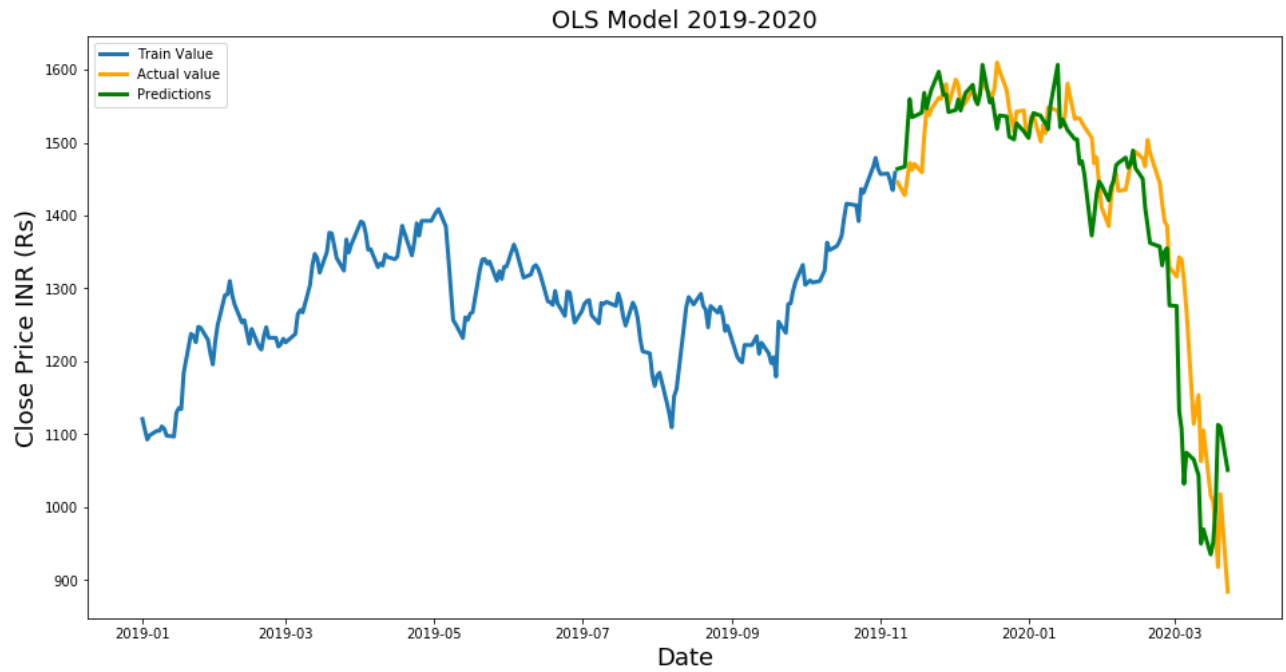


Fig. 15 : Reliance OLS Prediction for 2019-2020

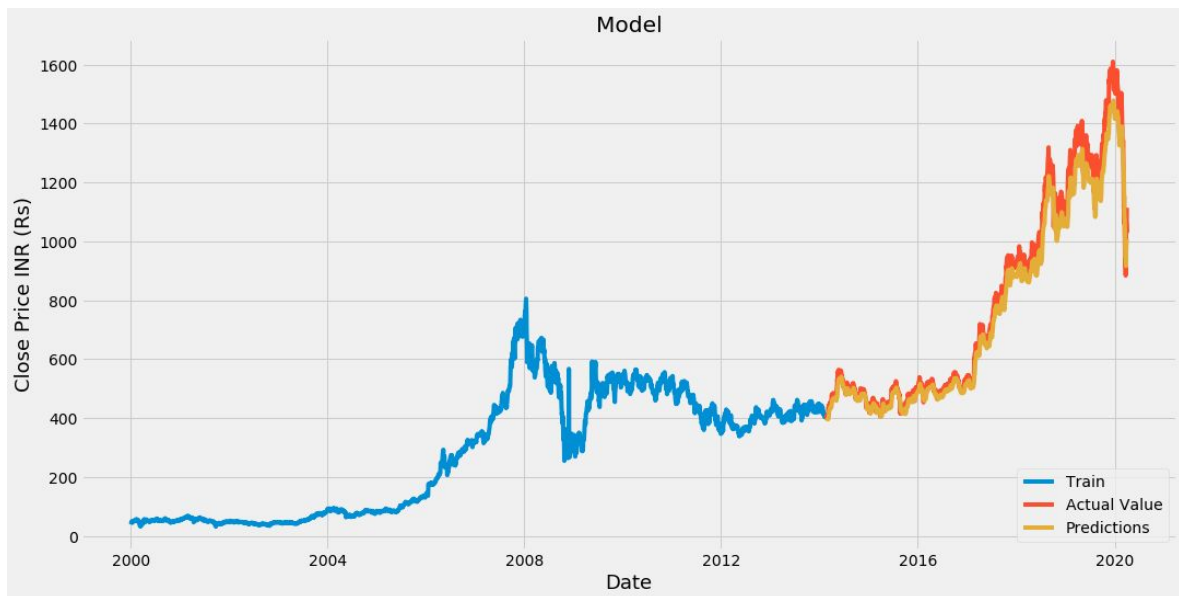


Fig. 15 : Reliance LSTM prediction for 2000-2020

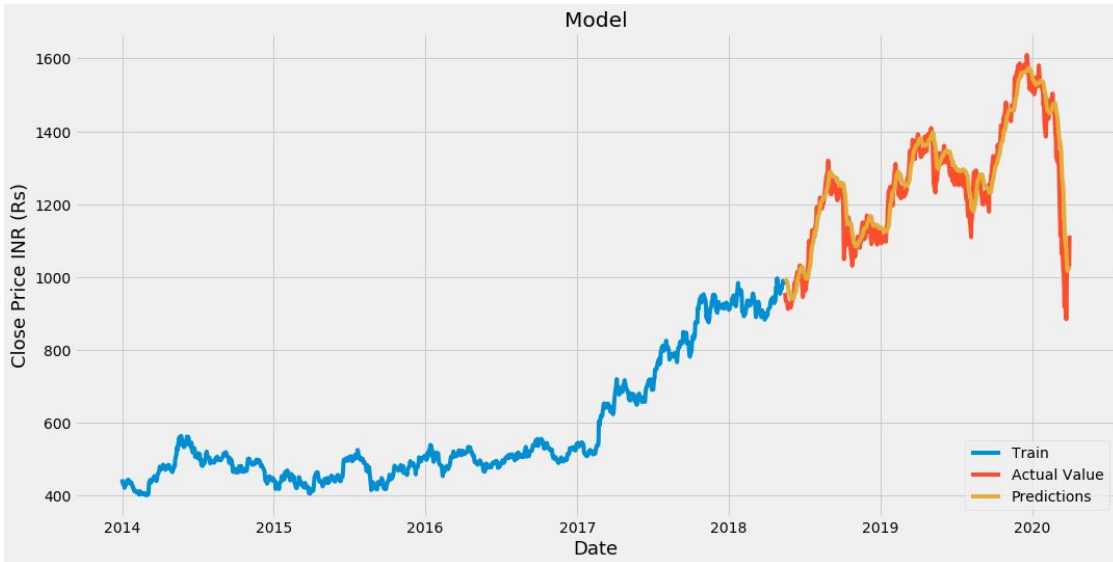


Fig. 16 : Reliance LSTM prediction for 2014-2020

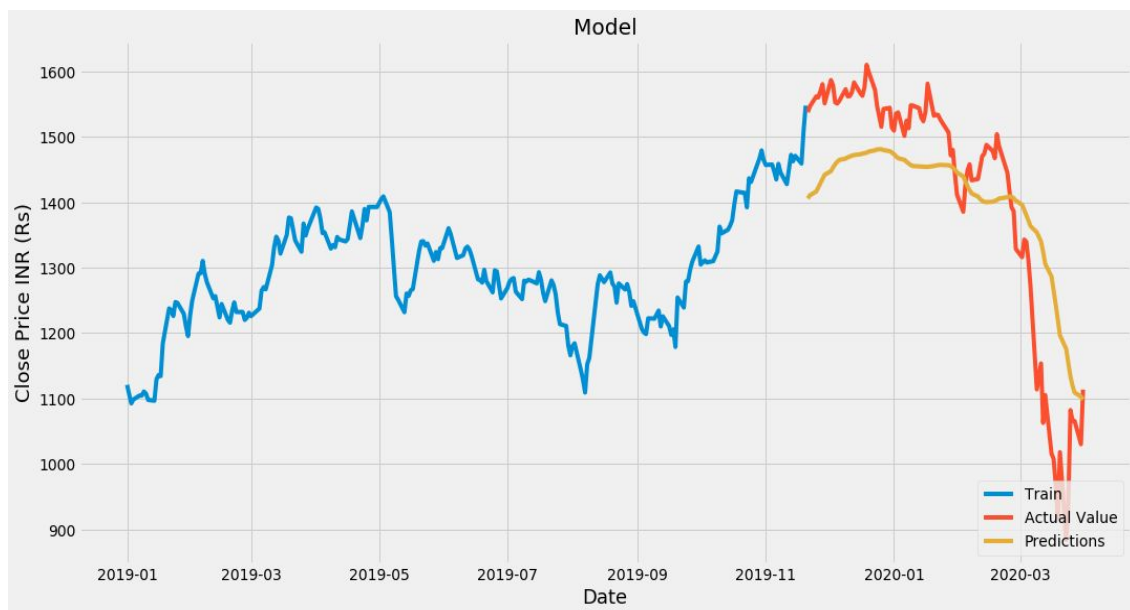


Fig.17 : Reliance LSTM prediction for 2019-2020



Fig.18 : Reliance SVR prediction for 2000-2020

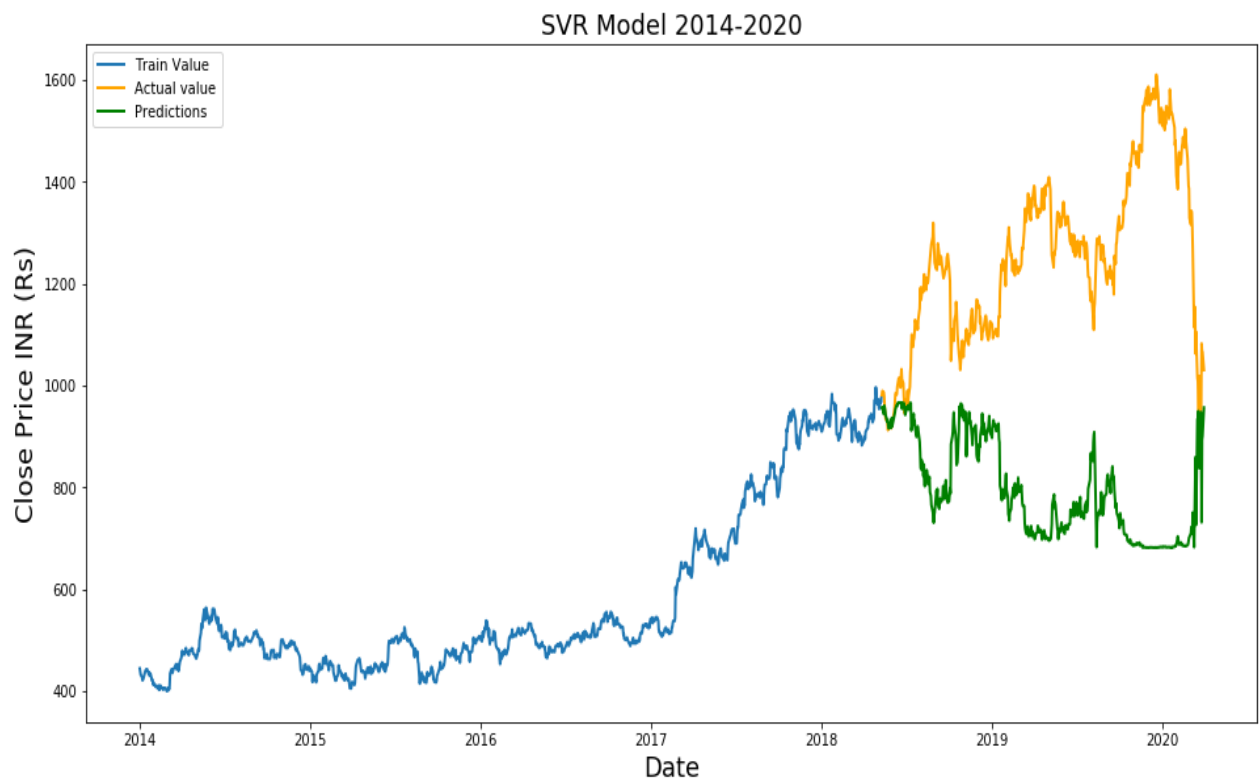


Fig.19 : Reliance SVR prediction for 2014-2020

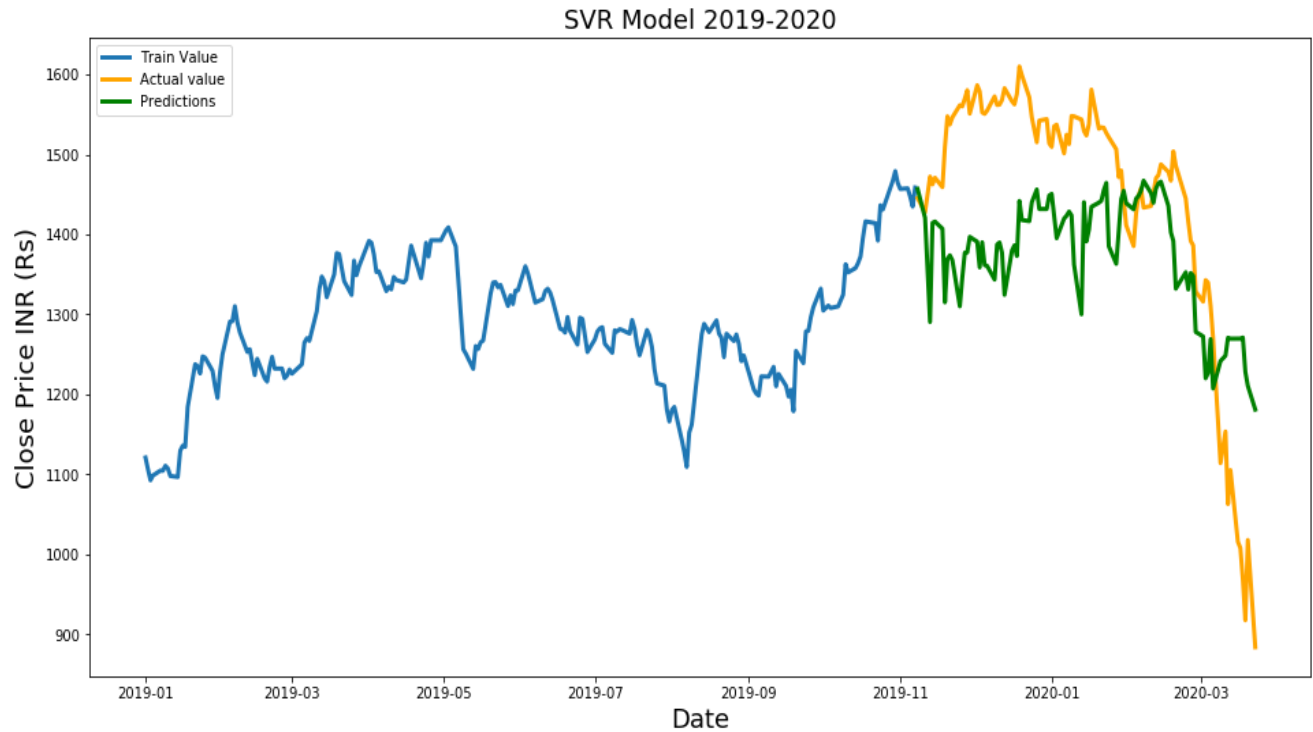


Fig.20 : Reliance SVR prediction 2019-2020

Table 4 : Reliance RMSE

Time Frame	OLS	SVR	LSTM
2000-2020	14.07	530.57	49.02
2014-2020	21.75	523.91	53.75
2019-2020	32.58	153.02	114.63

Table 5 : Reliance MAPE

Time Frame	OLS	SVR
2000-2020	0.0115	0.2560
2014-2020	0.013	0.2421
2019-2020	0.017	0.091

2) TCS - The Tata Consultancy Services Ltd. data also followed the same path. However, there exists one striking feature: SVR dropped its Mirror-image-forming pattern and followed the trend of the actual stock price for the timeframe of one year (Fig. :)

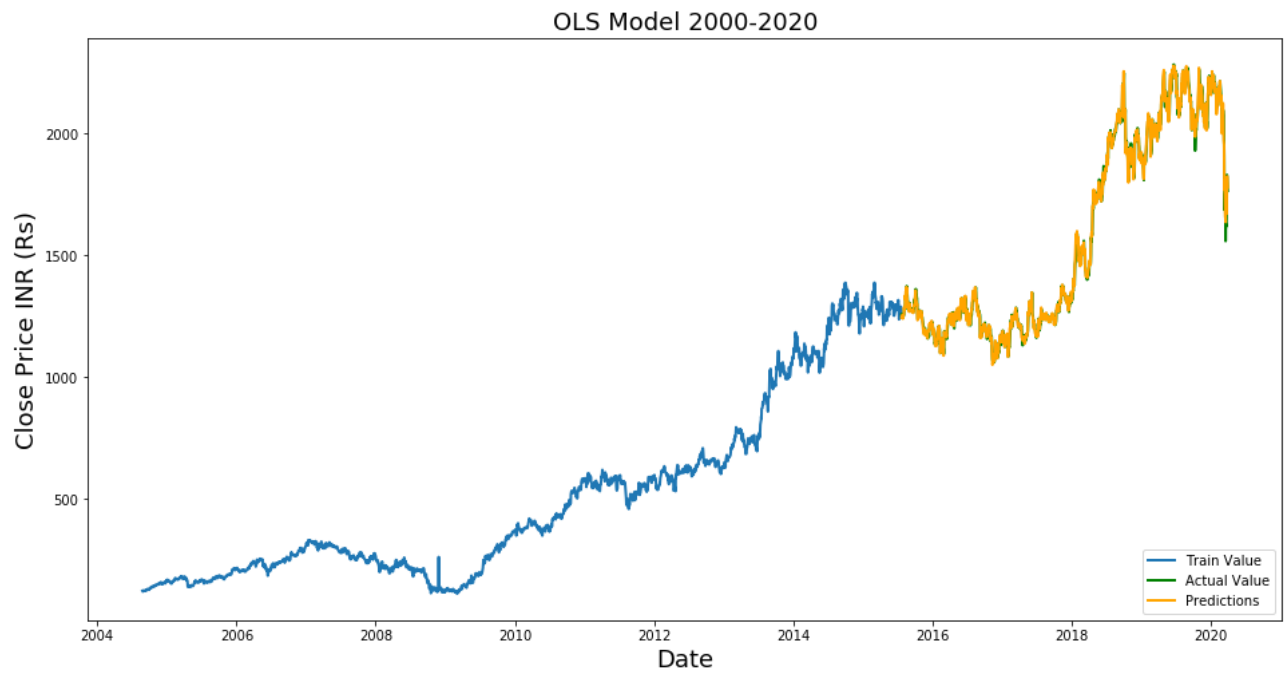


Fig.21 : TCS OLS Prediction for 2000-2020

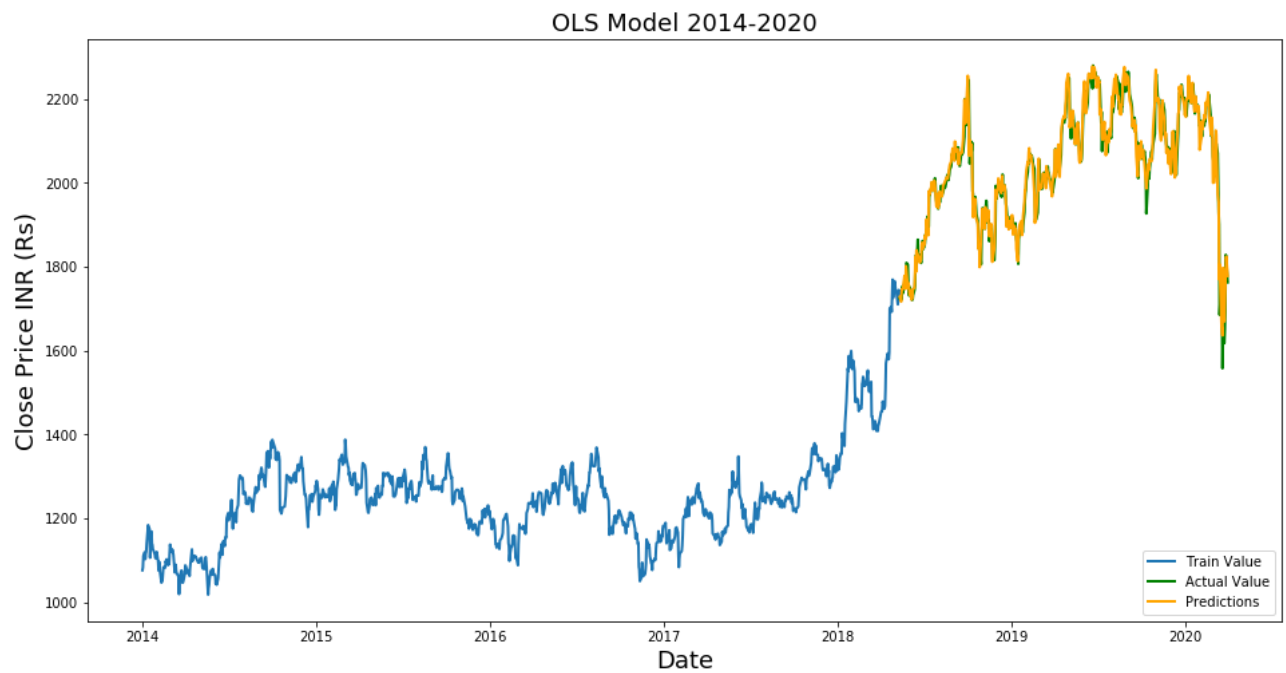


Fig.22 : TCS OLS Prediction 2014-2020

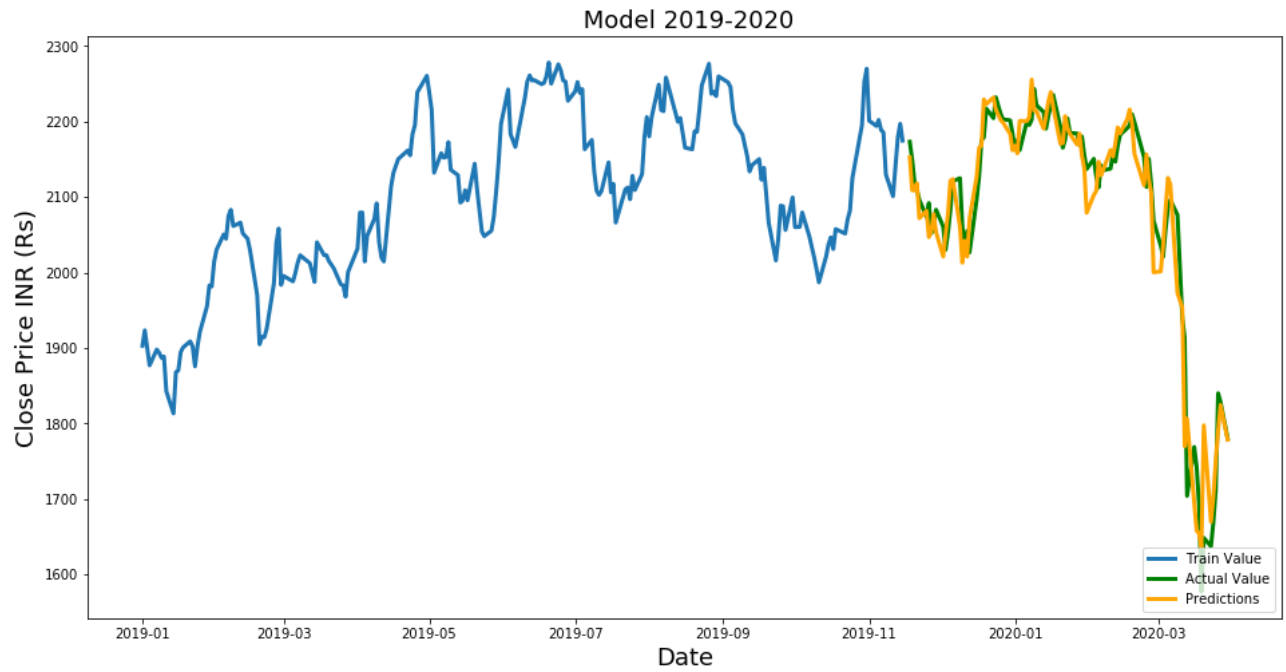


Fig. 23 : TCS OLS Prediction 2019-2020

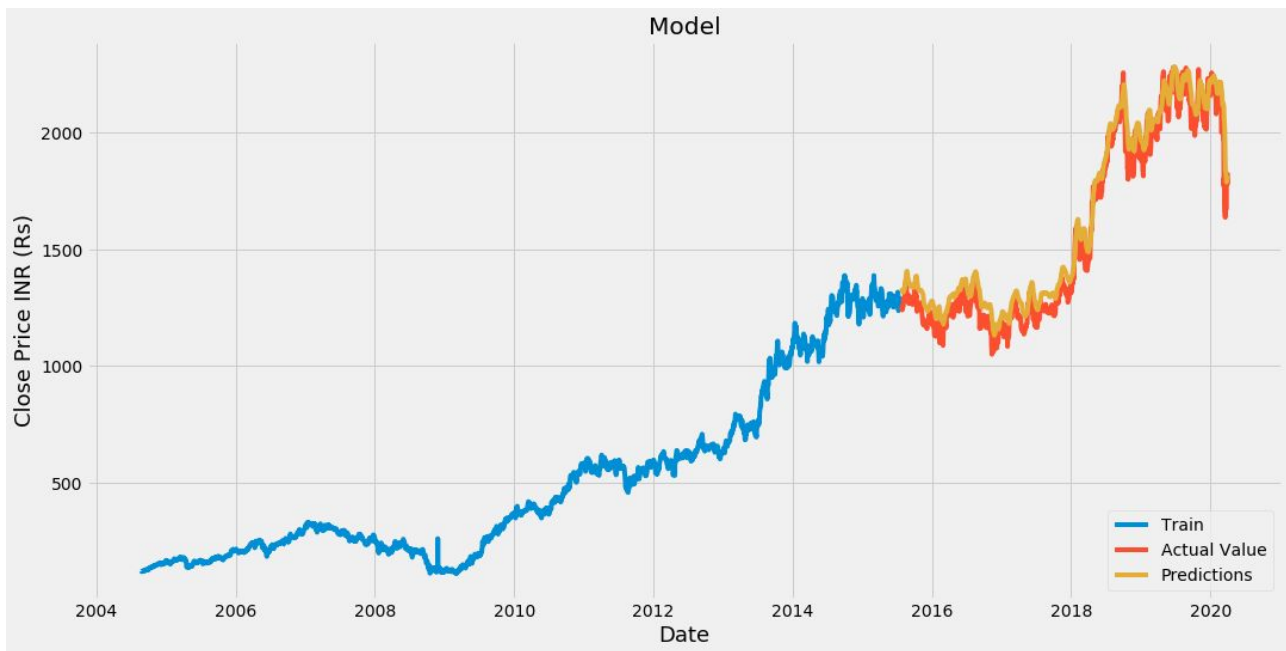


Fig.24 : TCS LSTM Prediction for 2000-2020

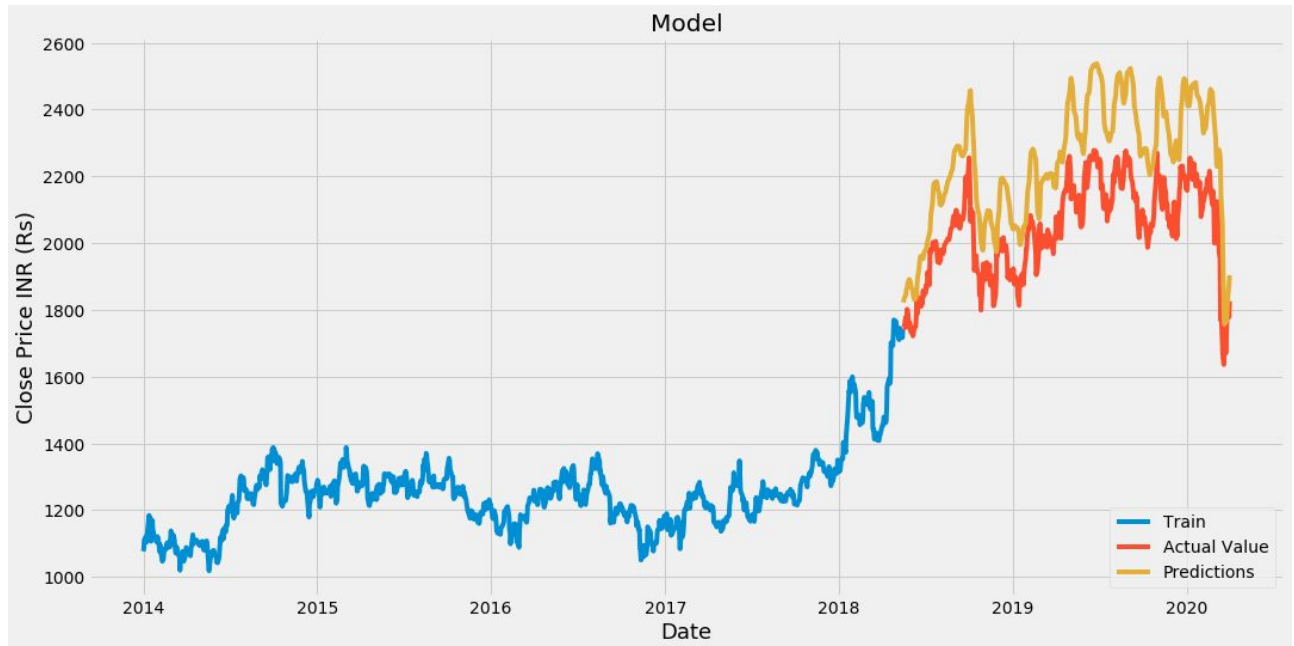


Fig.25 : TCS LSTM Prediction for 2014-2020

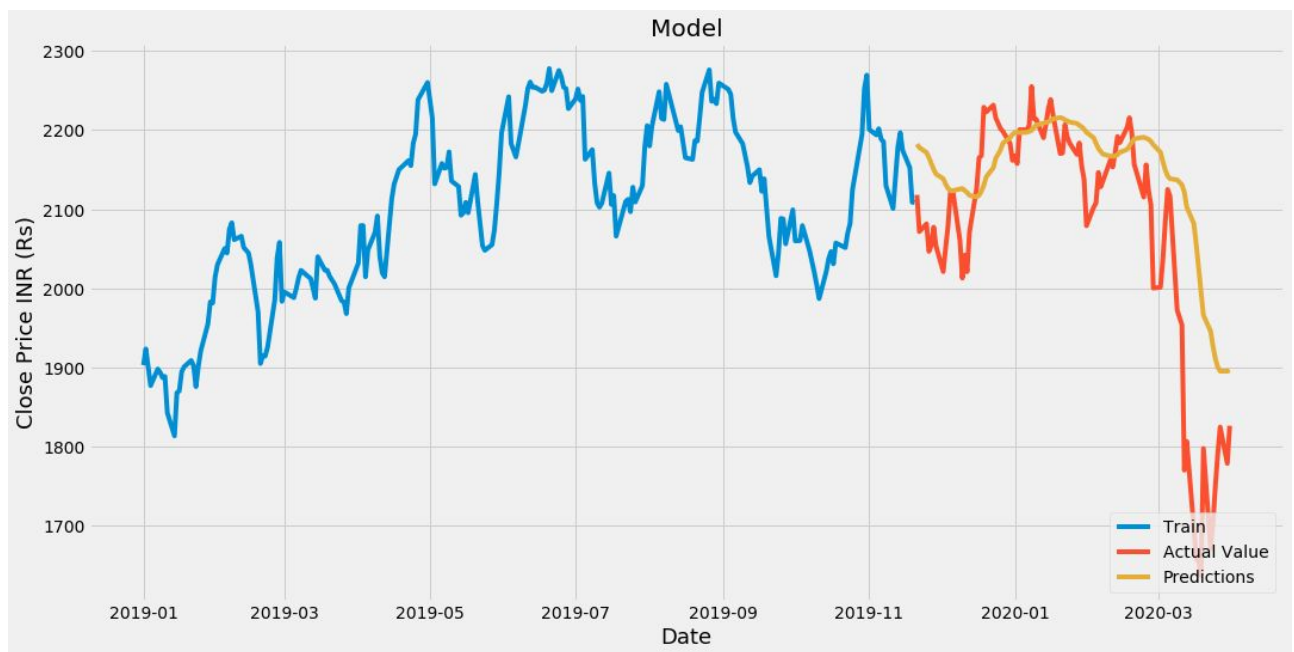


Fig.26 : TCS LSTM Prediction for 2019-2020

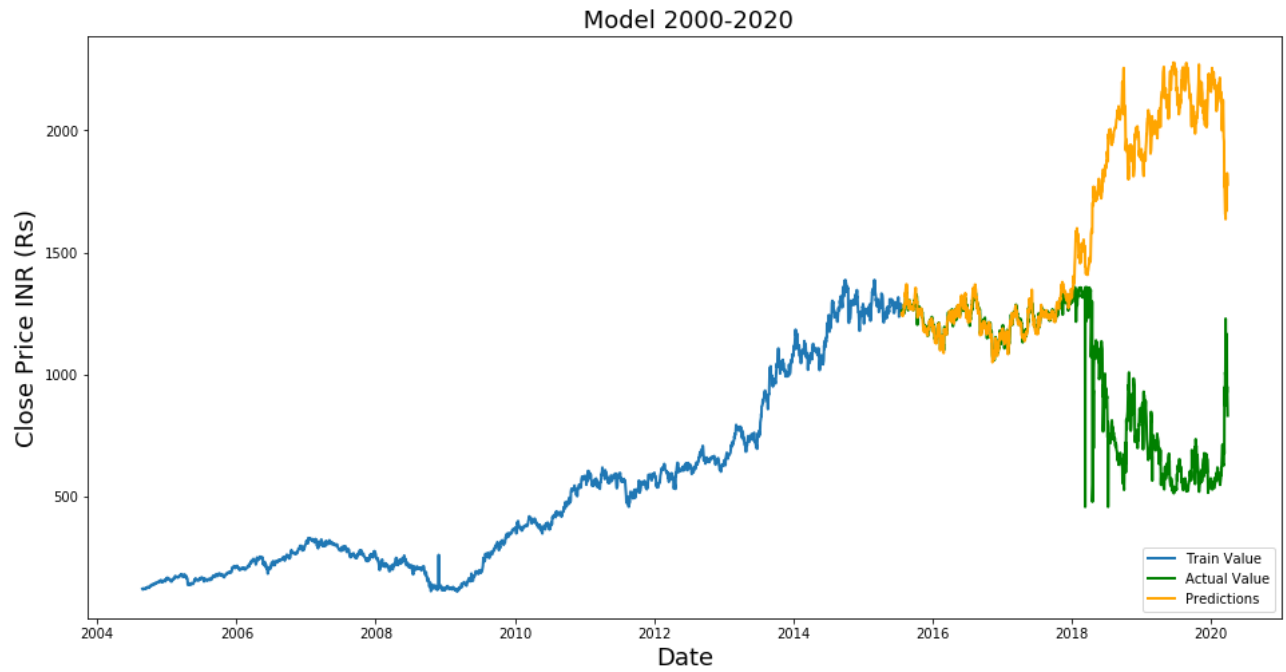


Fig.27 : TCS SVR prediction 2000-2020

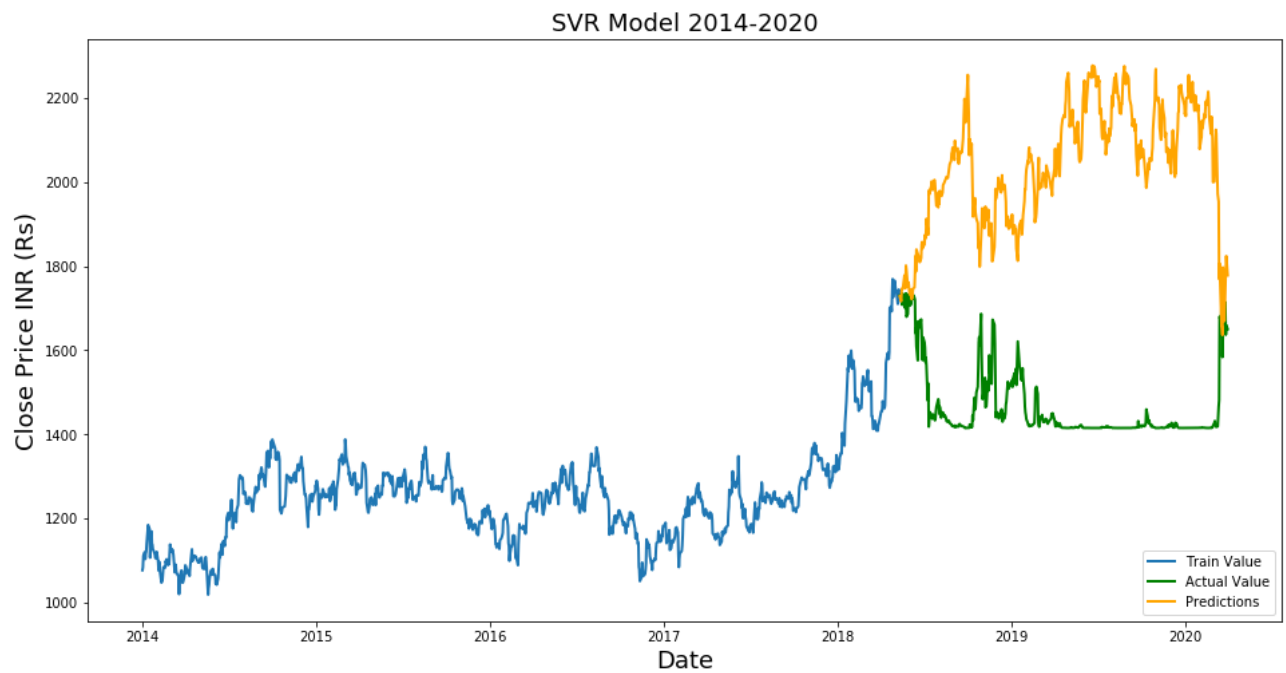


Fig. 28: TCS SVR prediction 2014-2020

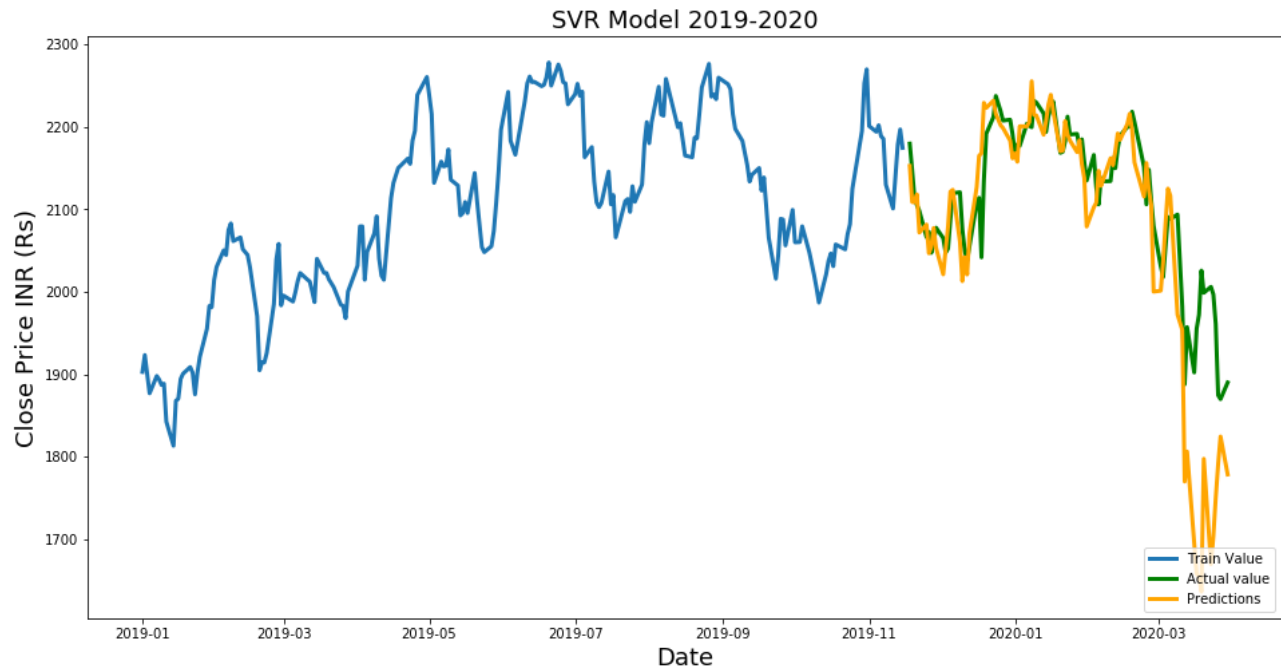


Fig.29 : TCS SVR prediction 2019-2020

Table 6 : TCS RMSE

Time Frame	OLS	SVR	LSTM
2000-2020	23.77	872.90	52.48
2014-2020	31.53	613.66	56.58
2019-2020	40.13	94.33	112.89

Table 7 : TCS MAPE

Time Frame	OLS	SVR
2000-2020	0.0105	0.2784
2014-2020	0.0117	0.2739
2019-2020	0.0149	0.028

3) Sun Pharmaceutical - The most surprising finding emerged out from the analysis of time-series data of Sun Pharmaceuticals. Though, both LSTM and OLS lost their efficiency

in forecasting as time frame became smaller and smaller but the loss was not of substantial nature, as can be seen from the graphs below. Additionally, the errors are also extremely low as compared to other companies.

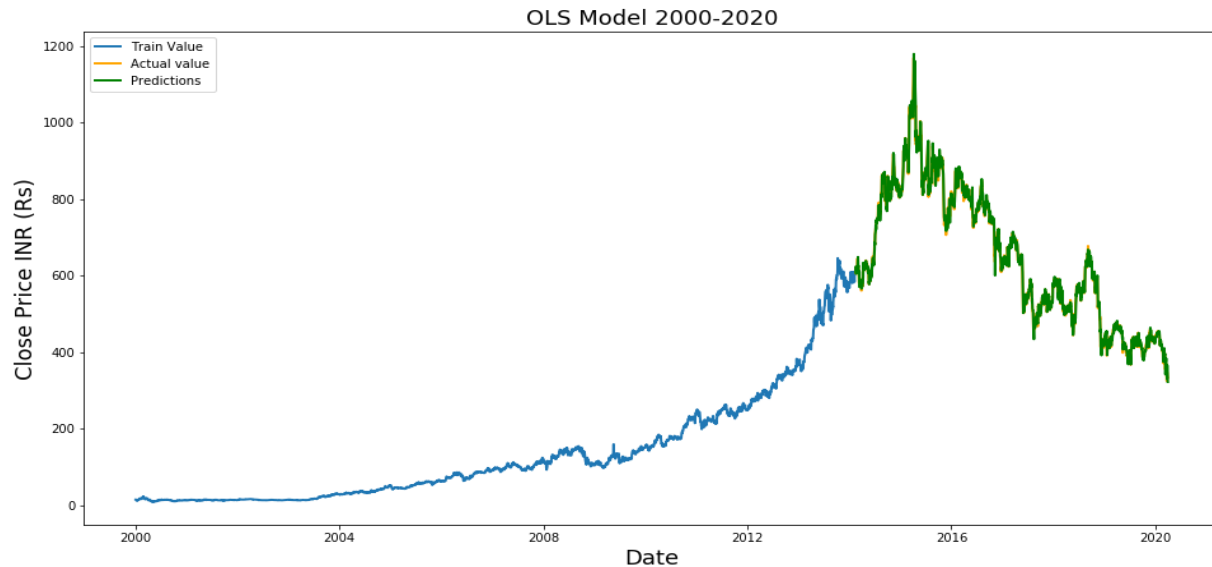


Fig.30 : Sun Pharmaceuticals OLS prediction 2000-2020

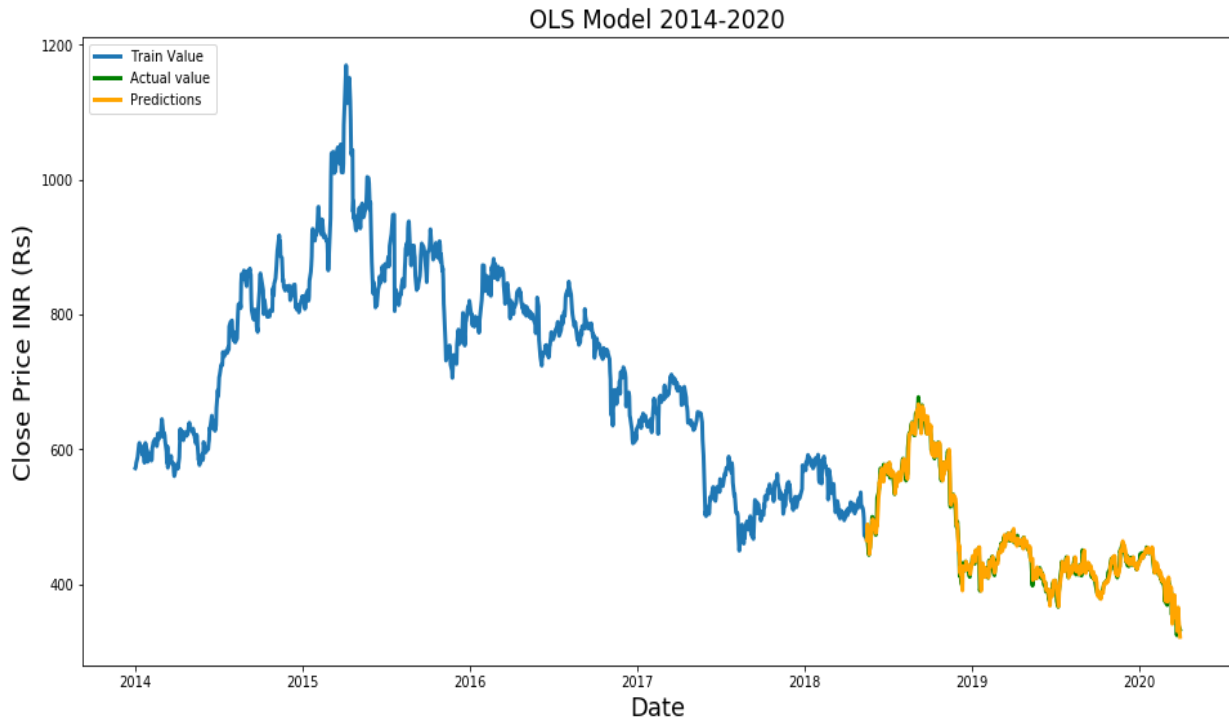


Fig. 31: Sun Pharmaceuticals OLS prediction 2014-2020

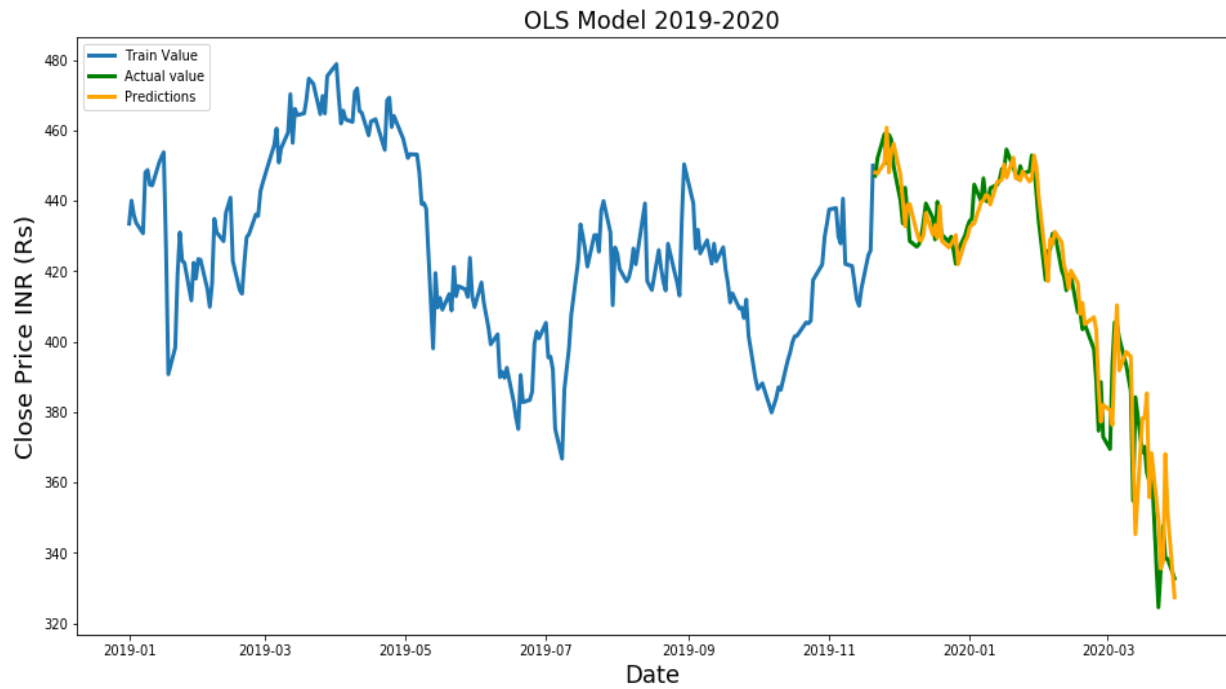


Fig. 32 : Sun Pharmaceuticals OLS prediction 2019-2020

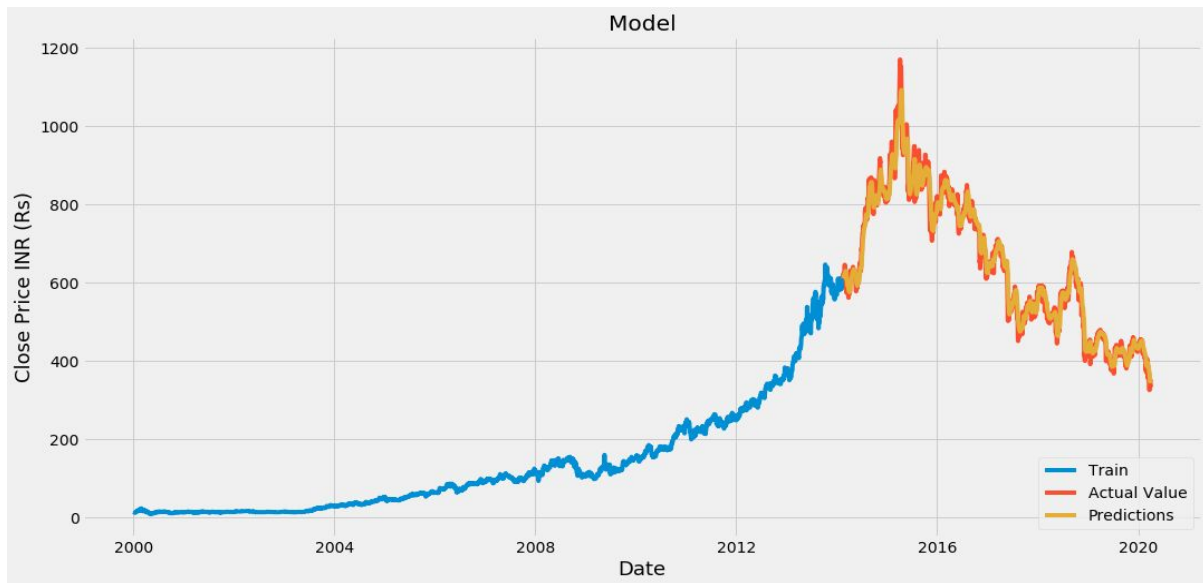


Fig. 33: Sun Pharmaceuticals LSTM prediction for 2000-2020

TM



Fig. 34: Sun Pharmaceuticals LSTM prediction for 2014-2020

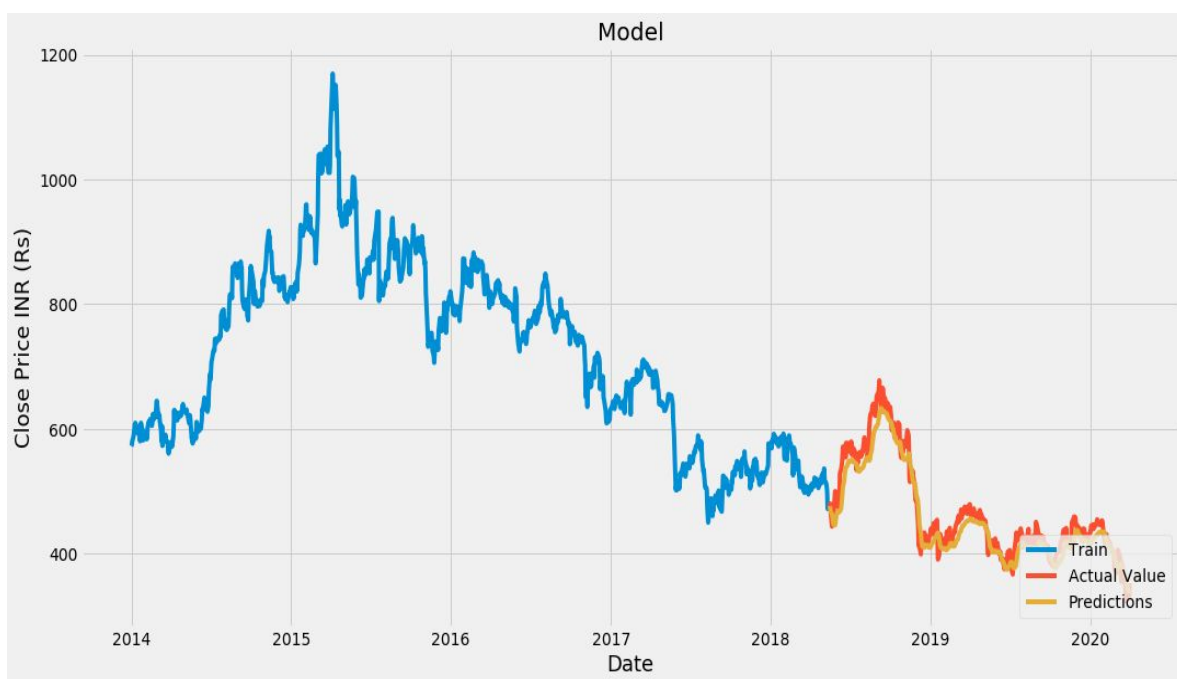


Fig. 35: Sun Pharmaceuticals LSTM prediction for 2019-2020

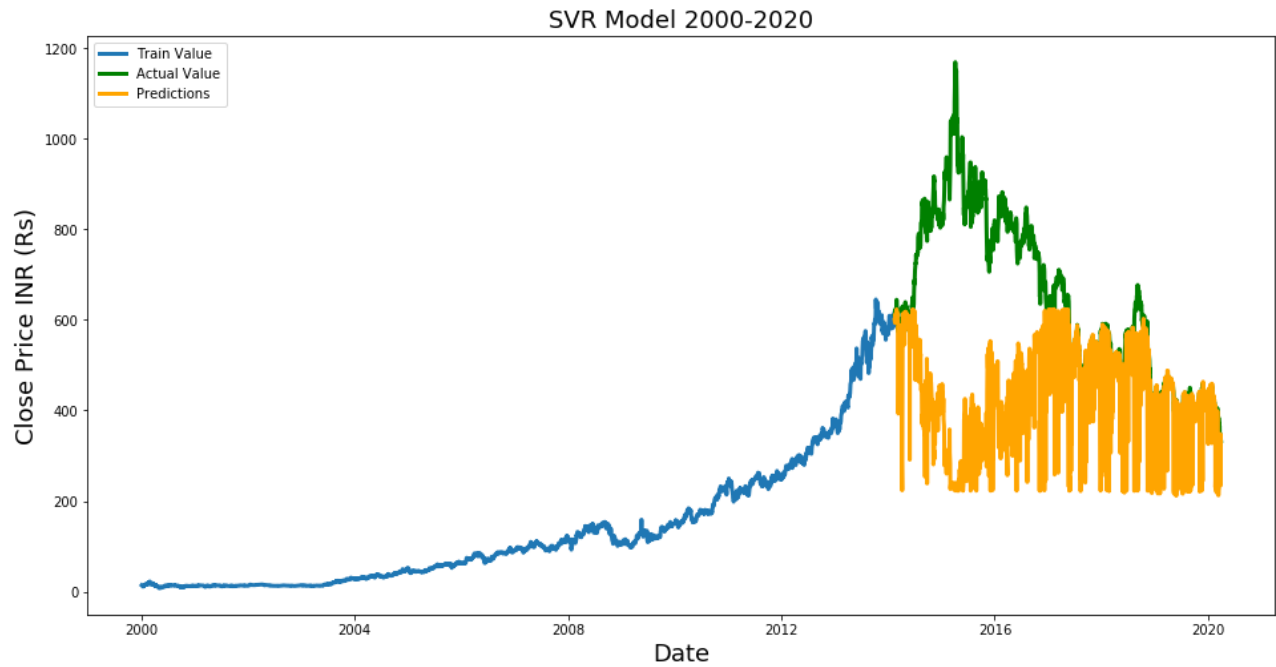


Fig.36 : Sun Pharmaceuticals SVR prediction for 2000-2020

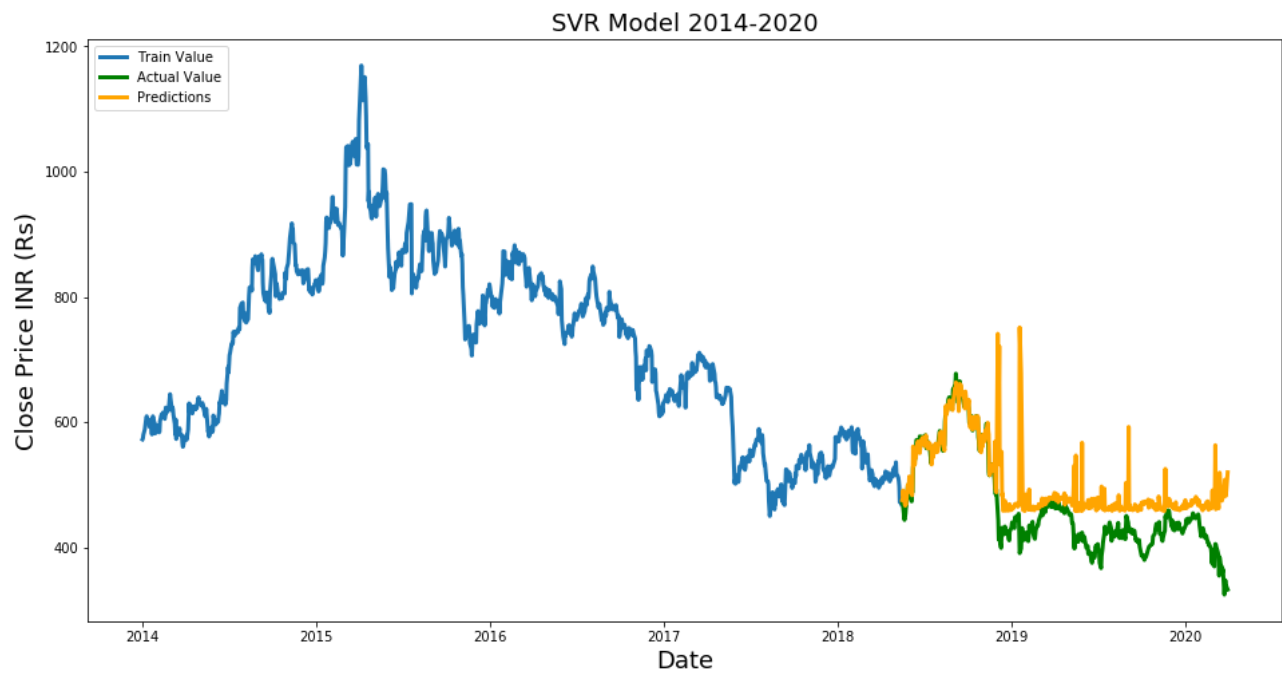


Fig.37 : Sun Pharmaceuticals SVR prediction for 2014-2020

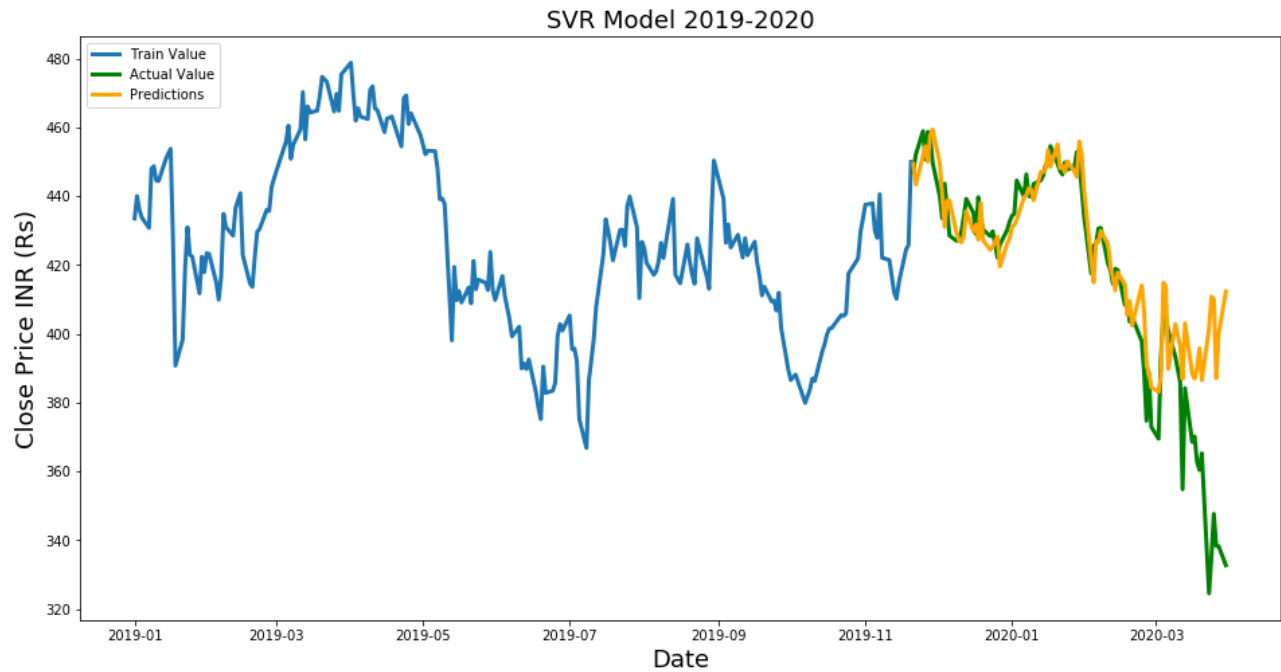


Fig. 38: Sun Pharmaceuticals SVR prediction for 2019-2020

Table 8: Sun Pharma RMSE

Time Frame	OLS	SVR	LSTM
2000-2020	12.09	312.61	24.61
2014-2020	9.92	59.73	21.96
2019-2020	9.054	20.08	20.70

Table 9: Sun Pharma MAPE

Time Frame	OLS	SVR
2000-2020	0.01396	0.2784
2014-2020	0.01576	0.9605
2019-2020	0.01629	0.0300

4) HDFC Bank -

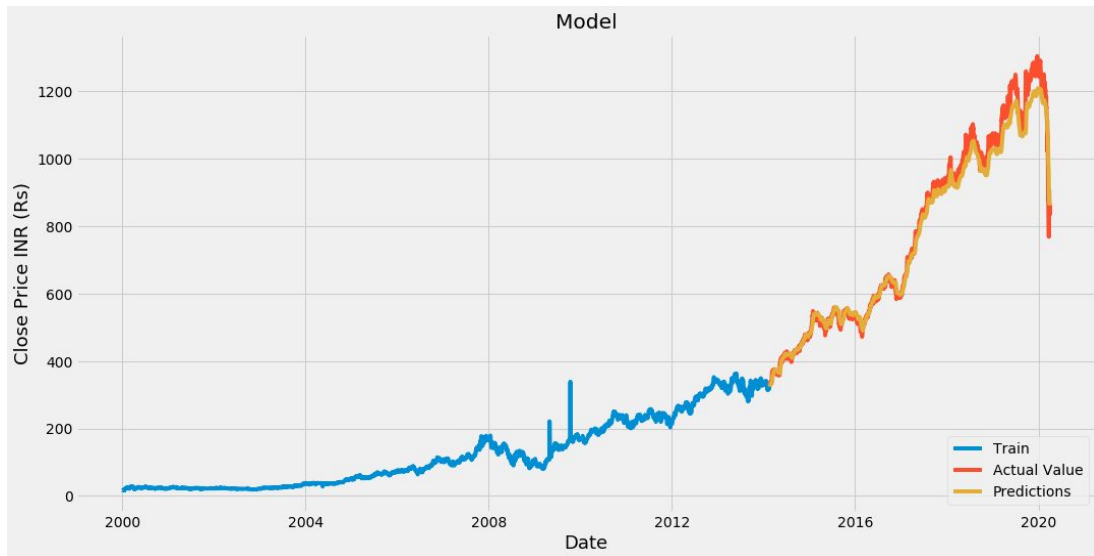


Fig.39 : HDFC Bank LSTM prediction for 2000-2020

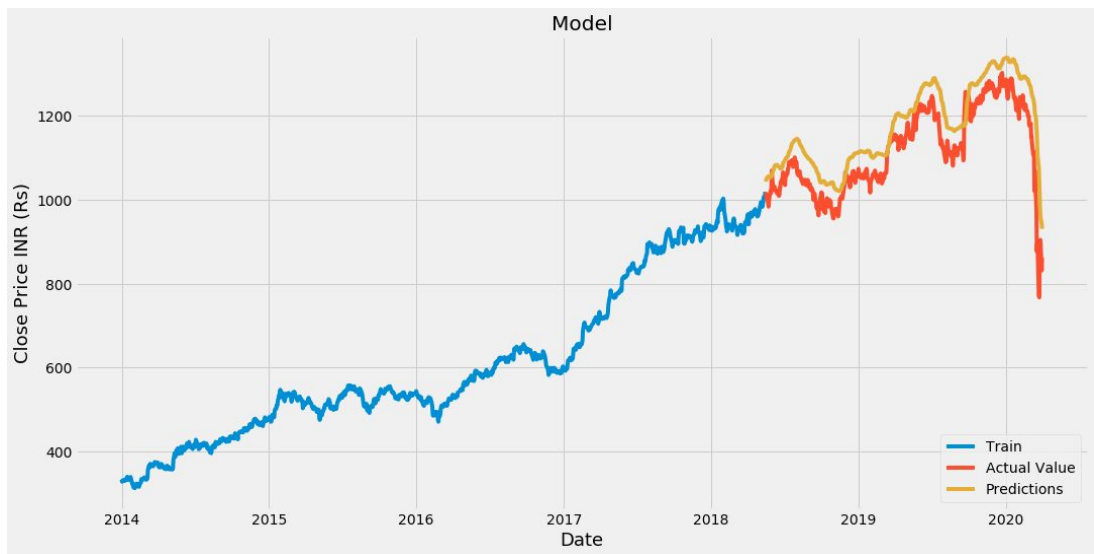


Fig.40 : HDFC Bank LSTM prediction for 2014-2020

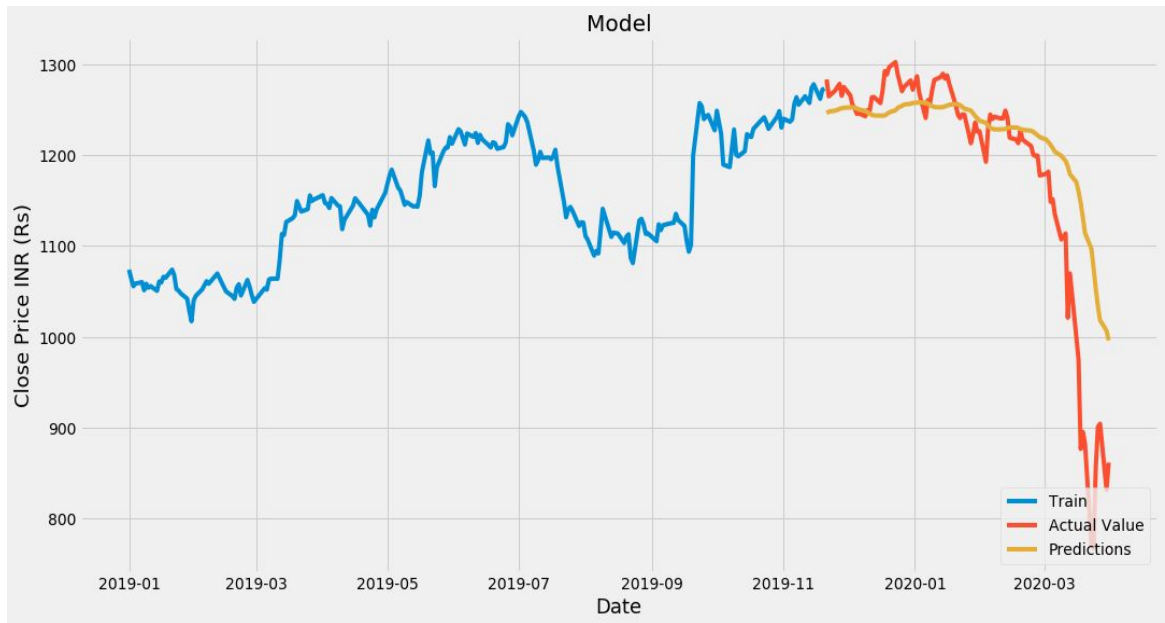


Fig.41 : HDFC Bank LSTM prediction for 2019-2020

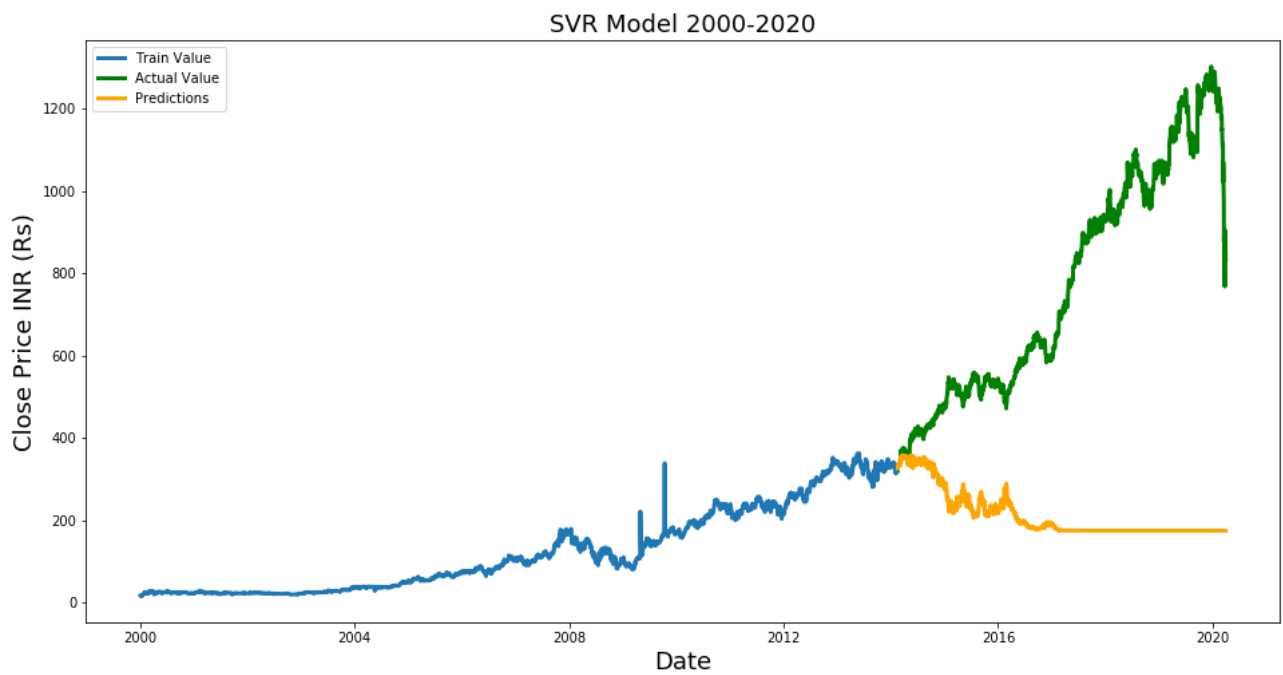


Fig. 42: HDFC Bank SVR prediction for 2000-2020

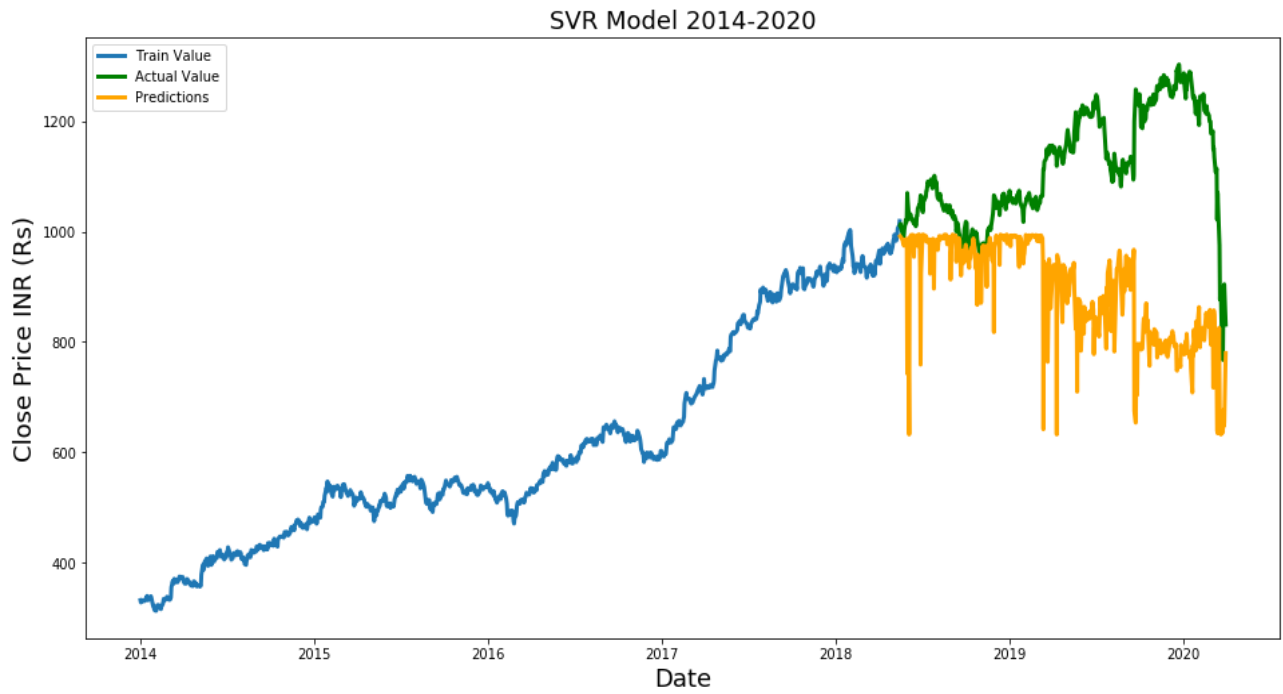


Fig.43 : HDFC Bank SVR prediction for 2014-2020

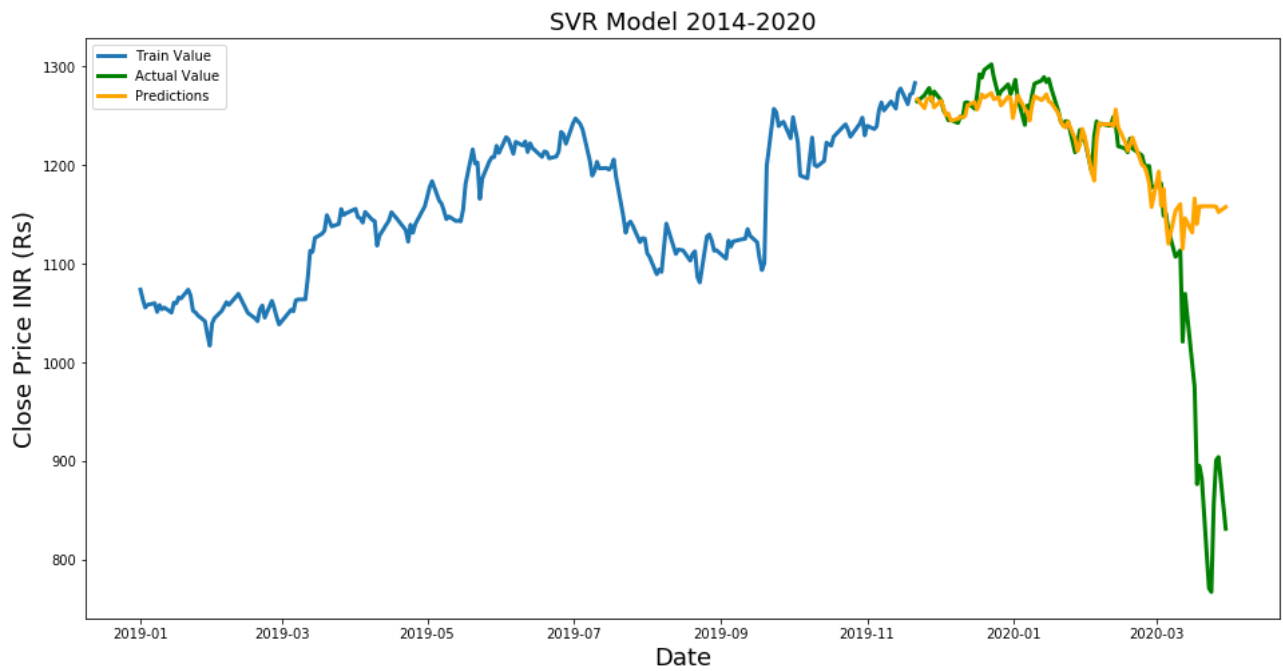


Fig. 44: HDFC Bank SVR prediction for 2019-2020

Table 10 : HDFC Bank RMSE

Time Frame	OLS	SVR	LSTM
2000-2020	9.76	648.24	32.94
2014-2020	15.12	279.13	68.18
2019-2020	24.55	102.28	85.83

Table 11: HDFC Bank MAPE

Time Frame	OLS	SVR
2000-2020	0.0079	0.6564
2014-2020	0.0089	0.1885
2019-2020	0.0141	0.0500

V. CONCLUSION :

The traditional technique of OLS turned out to be superior in our study, followed by the new school neural network technique of LSTM across all time frames. While, SVR yielded abysmally poor results, it is not a point of concern since it simply suggests that there exists a lack of non-linearities in our data frame. Thus, the old school machine learning SVR model appears to be incompetent as compared to the modern Deep learning model of LSTM based on multiple layers of networks. Literature suggests that the pharmaceutical industry gives good long-term returns and that could be the reason why our models performed exceptionally well for longer time frames in Sun Pharmaceutical's case.

Financial market analysis and forecasting continues to be a strenuous activity, constantly putting forth challenges with respect to drawing out relevant information from the data available and assessing its impact on market prices. Especially since we have trained and tested our model on historical data only, we should exercise precaution while extrapolating our findings for live-trading. Real-time trading involves more complexities pertaining to the very nature of the financial markets. Additionally, the behavioural aspect arising due to 'Fear and Greed' factor also adds to the already existing complications, which further questions the viability of stock prediction algorithms. The answer to these questions are further hazed out because of various stories on 'Fat finger errors' doing rounds in the stock market circuit. The Knight Capital Tragedy- an american company lost over \$460 million in a day back in 2012 due to a trading mistake committed by its automatically operated algorithm, which led to its acquisition later in 2013. Another story is of an Oil trader in futures market, Steve Perkins, who accidentally purchased 7 billion barrels (almost 69% of

total global trading) of crude oil in 2009 for almost £345 million using high-frequency algorithms while being under the influence of alcohol, and ended up losing his job (banned from trading). His company suffered a loss of £7.3 million.

The question, however, remains. And the only answer is to fabricate models that take into account these complexities and sentiments. Development of Sentiment analysis-based models, which make use of twitter and news feed for stock price forecasting, is one such step in this direction. These models, however, face a threat from fraudulent information. Therefore, we propose to research and construct a model which is hybrid in nature i.e. has features of both statistical and machine-learning models.

VI. REFERENCES :

- Cao, L., Tay, F. Financial Forecasting Using Support Vector Machines. *Neural Comput & Applic* 10, 184–192 (2001). <https://doi.org/10.1007/s005210170010>
- Fama, Eugene F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25, no. 2 (1970): 383–417. Accessed April 7, 2020. doi:10.2307/2325486.
- Bachelier, L. (1900). "Théorie de la spéculation". *Annales Scientifiques de l'École Normale Supérieure*. 17: 21–86. doi:10.24033/asens.476. ISSN 0012-9593
- Mandelbrot, Benoit (January 1963). "The Variation of Certain Speculative Prices". *The Journal of Business*. 36 (4): 394. doi:10.1086/294632. ISSN 0021-9398
- Samuelson, Paul A. (23 August 2015), "Proof that Properly Anticipated Prices Fluctuate Randomly", *The World Scientific Handbook of Futures Markets*, World Scientific Handbook in Financial Economics Series, 5, WORLD SCIENTIFIC, pp. 25–38, doi:10.1142/9789814566926_0002, ISBN 9789814566919
- Leigh, William, Cheryl J. Frohlich, Steven Hornik, Russell L. Purvis, and Tom L. Roberts. 2008. Trading with a Stock Chart Heuristic. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38: 93–104.
- Qiansheng Zhang, Jingru Zhang, Zisheng Chen, Miao Zhang, Songying Li. A New Stock Selection Model Based on Decision Tree C5.0 Algorithm. *Journal of Investment and Management*. Vol. 7, No. 4, 2018, pp. 117-124. doi: 10.11648/j.jim.20180704.12
- Bhuriya, Dinesh, Girish Kausha, Ashish Sharma, and Upendra Singh. 2017. Stock Market Prediction Using a Linear Regression. Paper presented at the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, April 20–22; vol. 2.
- Roondiwala, Murtaza, Harshal Patel, and Shraddha Varma. 2017. Predicting Stock Prices Using Lstm. *International Journal of Science and Research (IJSR)* 6: 1754–56.