Time-series Prediction of Stock Data
Machine Learning Fall 2018 Project

Reet Barik and Sriyandass Adidass

December 14, 2018

Abstract

What started off as an exercise to predict the highs and lows of stock data given external stimuli of political and/or economical nature, the objective of our project evolved to that of proving the effect of NIFTY as an exogenous variable to enhance the prediction of stock data. Multiple models were used to test this hypothesis and the results obtained have been described in detail in this report.

1 Introduction

Times-series prediction is an unsolved problem and it has found the most application in the field of finance and economics. But there has been no model developed with enough accuracy that can drive investors to consider them as a real-life stock predictor. So much so that any novice investor can outperform the most sophisticated of the available stock prediction models. This is so because, to predict stock value accurately, a model needs a whole set of exogenous variables. The problem is that to make the exogenous variable set exhaustive, we risk adding an impractical amount of features. Moreover, this exercise is also futile in the sense that obtaining data for most of those exogenous variables is not possible given their non-quantifiable nature. To overcome these challenges, this project considers the NIFTY-50 index as a substitute for many such influential exogenous variables to enhance the prediction of stock data.
NIFTY-50 is the Indian Equity Markets benchmark index. It represents the weighted average of representative 50 Indian company stocks in 12 sectors (equivalent to NASDAQ in the United States). Due to the time constraints associated with this project, the scope has been limited to the Information Technology sector. This is so because, the IT sector being the major driving force behind the Indian economy in the past couple of decades, will act as a good approximate representation of all sectors and hence the proof/disproof of our hypothesis will hopefully translate to all other sectors constituting the Indian economy.

2 Problem Setup

The problem setup has been effectively summarized as follows:

- Use univariate models as analytical tools to figure out parameters like order of seasonality in the data etc.
- Predict NIFTY with reasonable accuracy using various multivariate models.
- Predict stock for each company being considered without NIFTY as an exogenous variable.
- Predict stock for each company being considered with NIFTY as an exogenous variable.
- Compare the performance of the models with and without NIFTY and document the result that prove/disprove our hypothesis.

3 Solution Approach

This section attempts to explain the assumptions behind the decisions taken for the experimental setup. They are as follows:

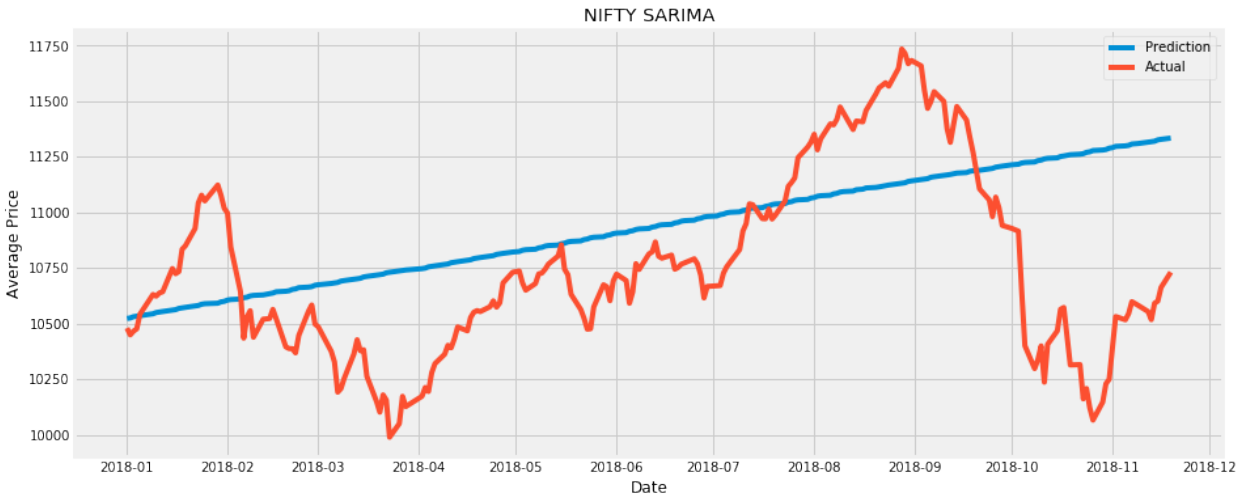
- NIFTY represents the weighted average of representative 50 Indian company stocks in 12 sectors and hence, it is intuitively a good representation of the set of exogenous variables to the models predicting the average stock price of different companies.
- There is a significant correlation between NIFTY and the IT sector stocks. Hence, the top five in the terms of the stocks that have been invested in the most have been chosen to be the best approximation of the whole sector. These five companies are Tata Consultancy Services (TCS), Infosys (INFY), Tech Mahindra (TECHM), HCL Technologies (HCLTECH), and Wipro (WIPRO).
- Two univariate models have been used here, not to be effectively used to prove/disprove our hypothesis, but as analytical tools to understand the underlying structure and seasonality of the data. They are as follows:
  1. SARIMAX(Seasonal Autoregressive Integrated Moving Average) [1]
  2. Holt-Winters [2]
- Four multivariate models have been used to:
  - train on and predict the NIFTY average price
  - train on and predict the average stock price of the above mentioned companies once without NIFTY as an exogenous variable and once with.

  The models used are as follows:
  1. VAR (Vector Autoregression) [3]
  2. XGBOOST [4]
  3. LSTM (Long Short-Term Memory) [5]
  4. GRU (Gated Recurrent Unit) [6]
- Compare all the results to finally accept or reject our hypothesis.

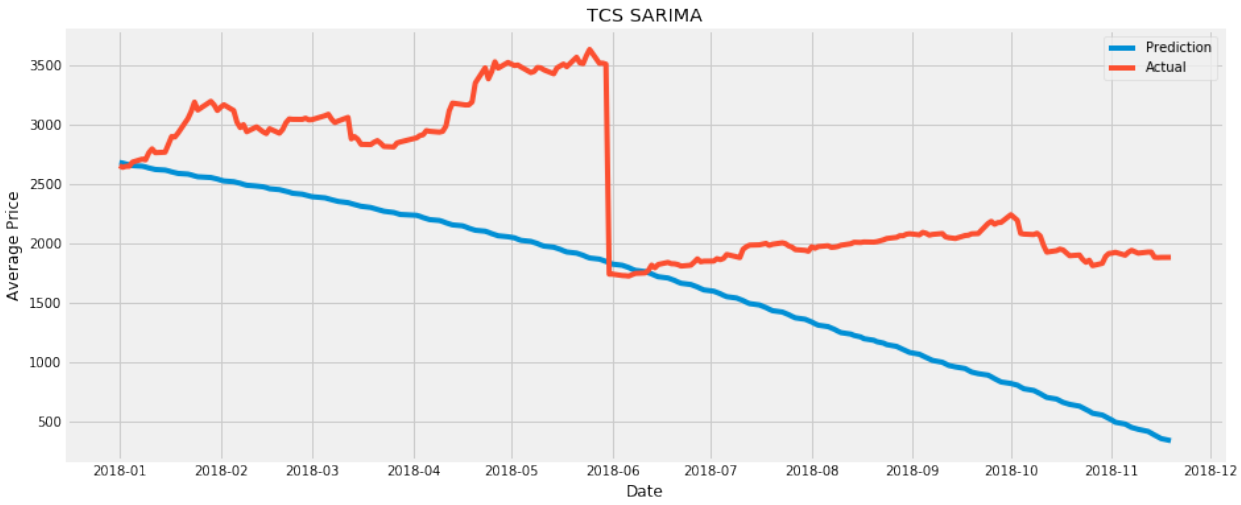
4 Experiments and Results

- **Data** used in our experiments is explained as follows:
  - The NIFTY data collected spans from 1 January 2013 to the date of data collection which was 19 November 2018. The models were trained on the data from the year 2013-2017 and were subsequently tested on the data from the year 2018.
    Data source: NIFTY-50 Data
  - The IT Stock data collected spans from 1 January 2013 to the date of data collection which was 19 November 2018. The models were trained on the data from the year 2013-2017 and were subsequently tested on the data from the year 2018.
    Data source: IT Stock Data
- **Performance Metrics:**
  - MAPE (Mean Absolute Percentage Error): Since the objective is to drive investment based on the results obtained from the models' predictions, MAPE is a good performance metric since it is an encapsulation the percentage error of each predicted point.
  - RMSE (Root-mean-square Error): This metric is chosen because it captures variation of prediction with respect to the actual very well. As can be seen from the experiments, RMSE might be high even for small MAPE.
- **Pre-analysis:**
  - SARIMAX:

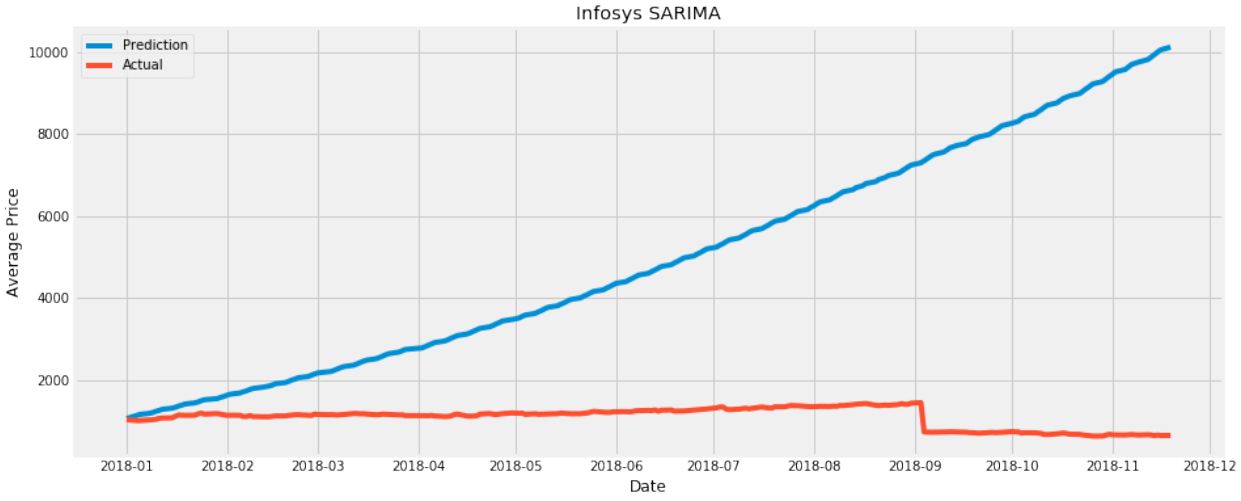
    The NIFTY prediction using SARIMAX is shown in the following figure:
    RMSE: 433.5275246569258
    MAPE: 3.2385394716378206



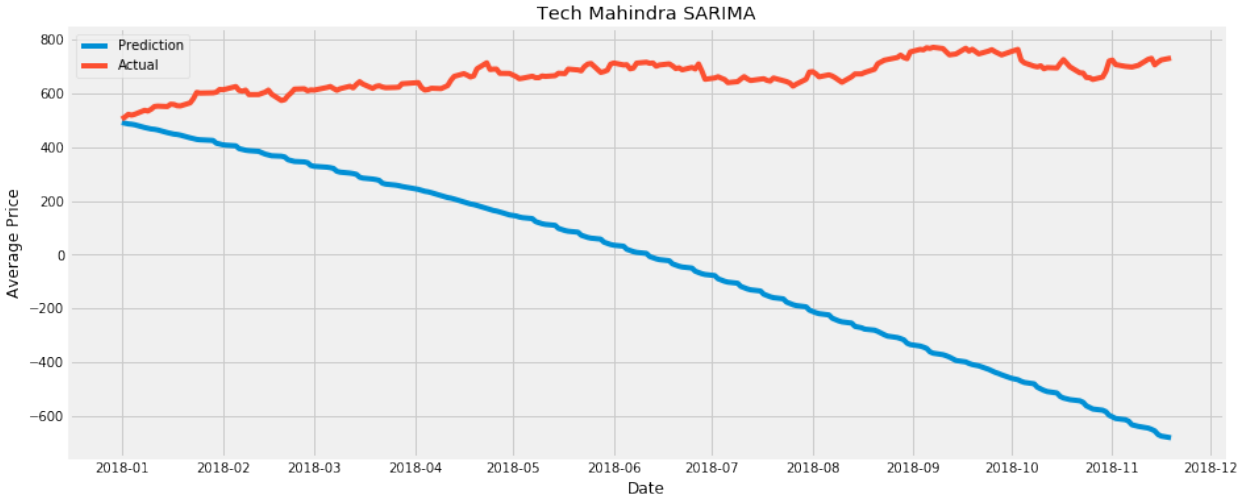
TCS stock prediction using SARIMAX is shown in the following figure:
RMSE: 945.9717331125928
MAPE: 33.63194540767255



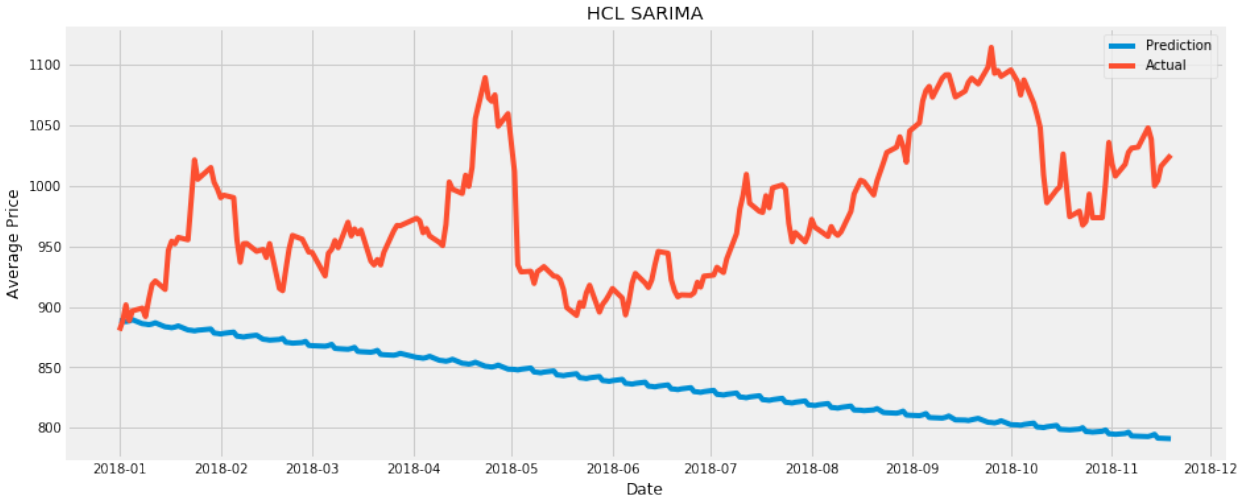
Infosys stock prediction using SARIMAX is shown in the following figure:
RMSE: 4716.794452598775
MAPE: 421.6301079069972



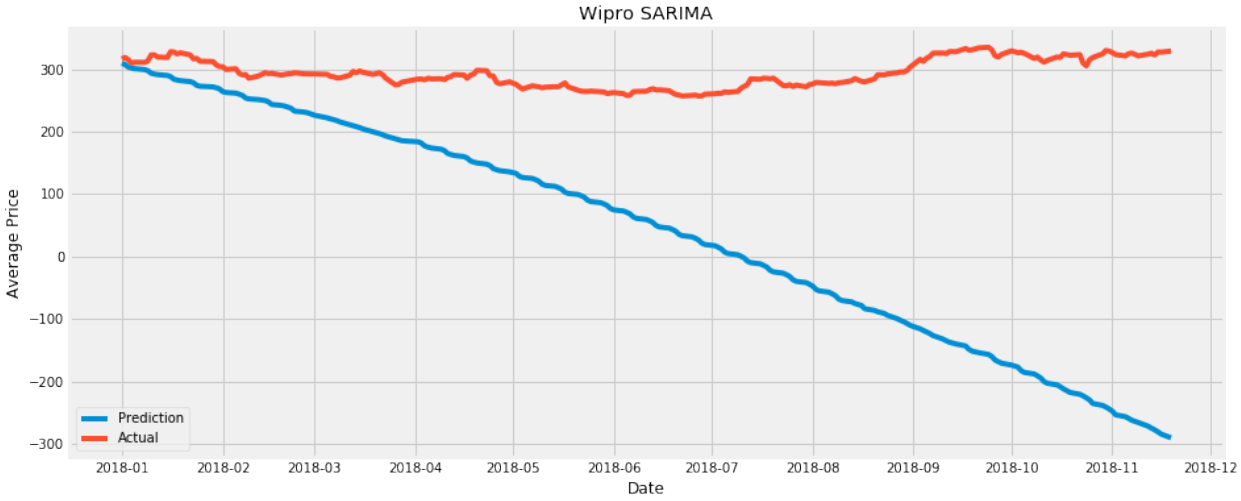
TECHM stock prediction using SARIMAX is shown in the following figure:
RMSE: 792.0110961669409
MAPE: 100.66066113767653



HCL stock prediction using SARIMAX is shown in the following figure:
RMSE: 158.53918494577724
MAPE: 13.906386560433429

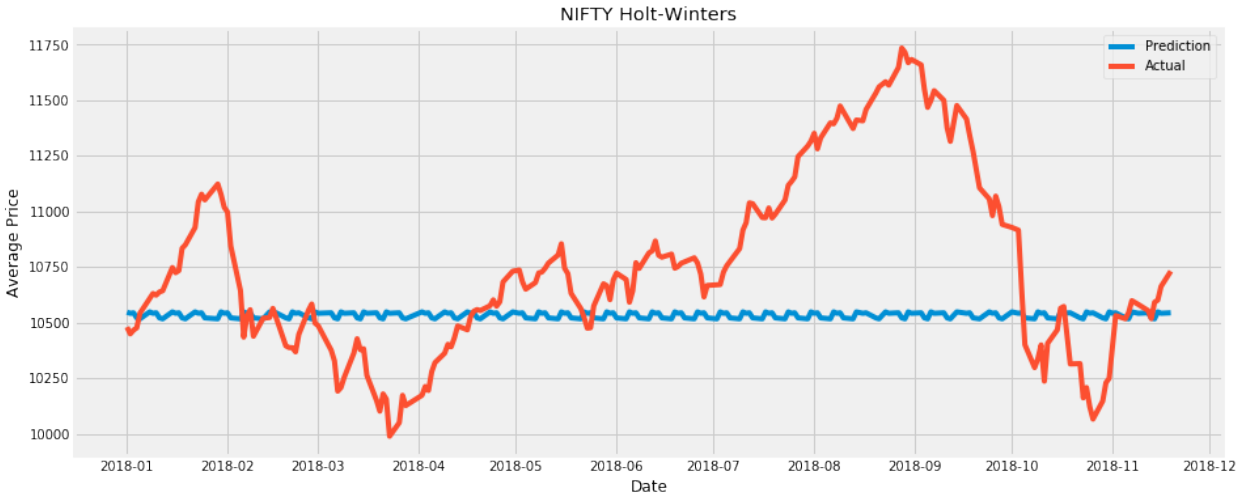


WIPRO stock prediction using SARIMAX is shown in the following figure:
RMSE: 309.2382668766522
MAPE: 83.94840932115724

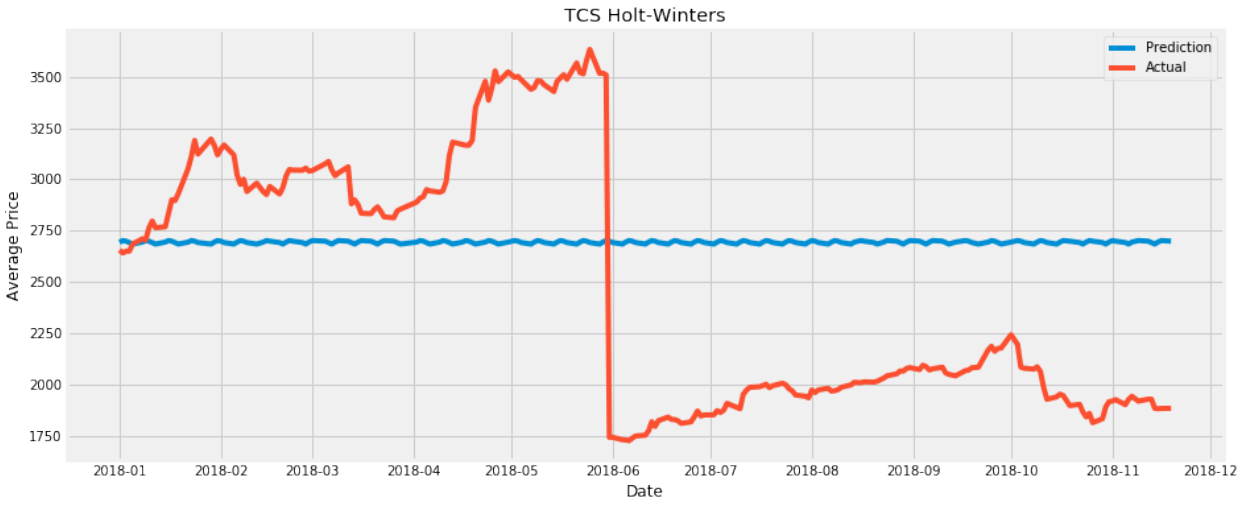


– Holt-Winters:

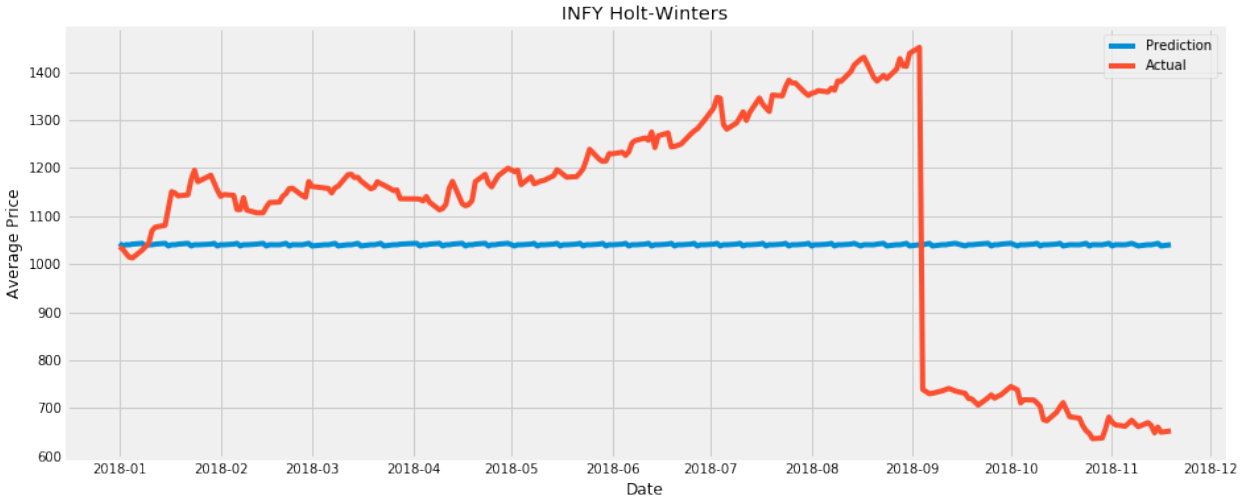
The NIFTY prediction using Holt-Winters is shown in the following figure:
RMSE: 455.7077009218795
MAPE: 3.0587489308351454



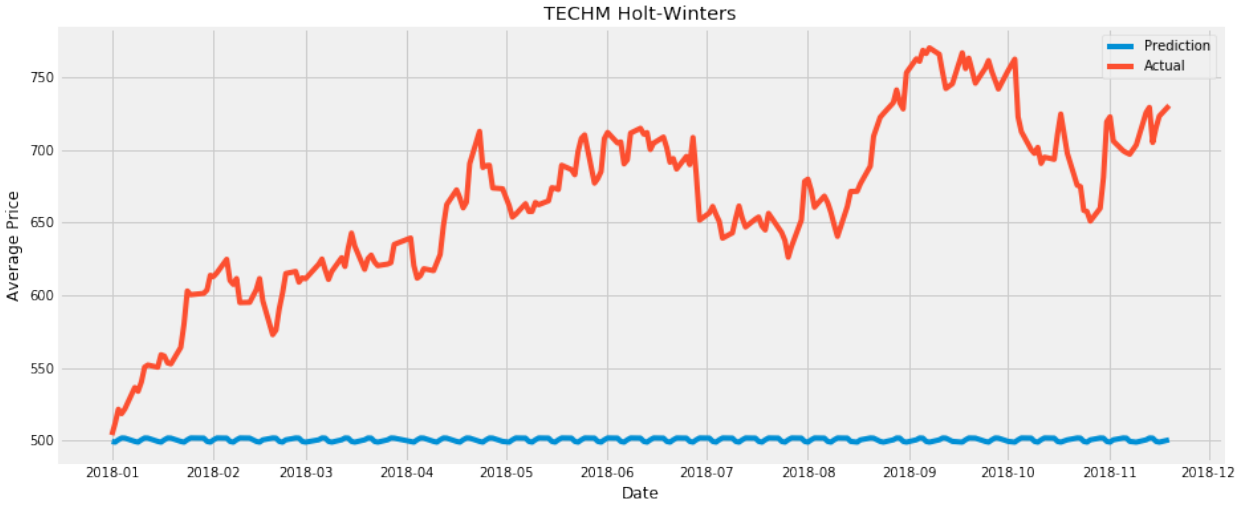
TCS stock prediction using Holt-Winters is shown in the following figure:
RMSE: 641.0020335147968
MAPE: 26.320737741126944



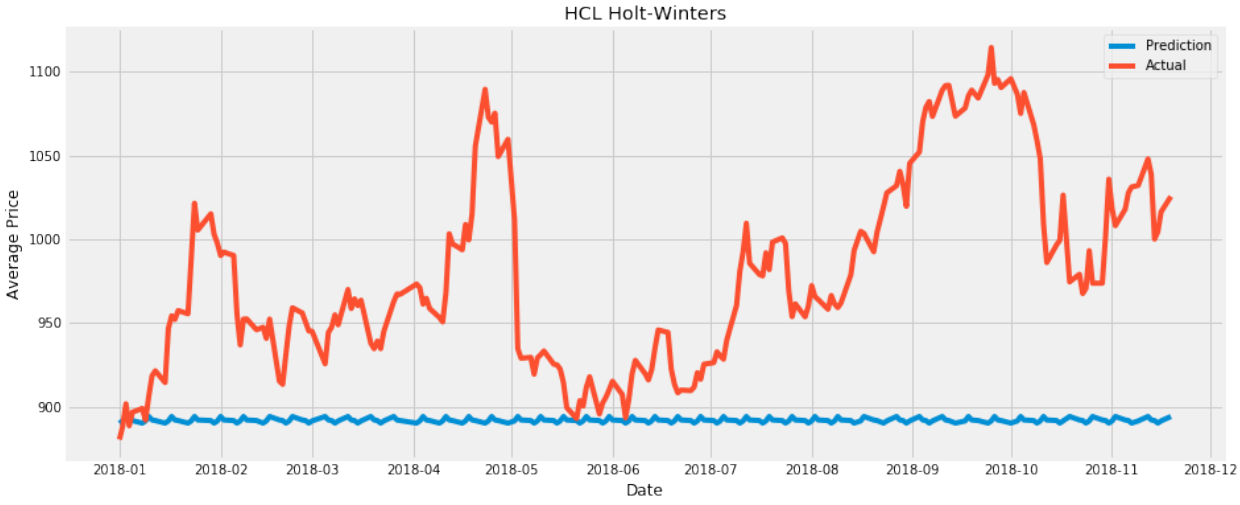
Infosys stock prediction using Holt-Winters is shown in the following figure:
RMSE: 245.3978697401031
MAPE: 22.341646080440093



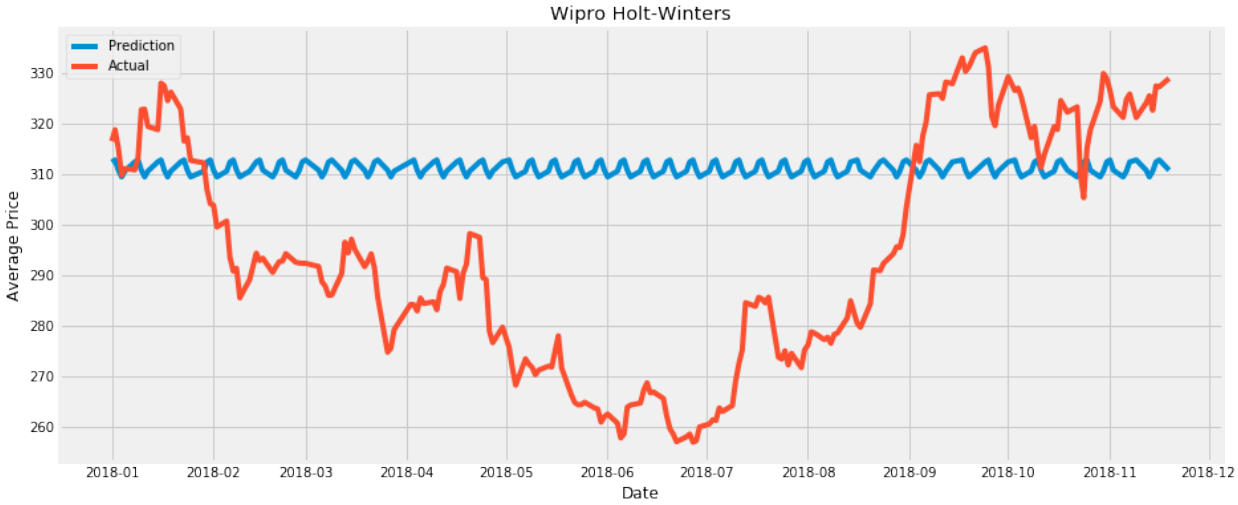
TECHM stock prediction using Holt-Winters is shown in the following figure:
RMSE: 172.3766079323786
MAPE: 23.91833864638295



HCL stock prediction using Holt-Winters is shown in the following figure:
RMSE: 102.04930035387645
MAPE: 8.429397185125426



WIPRO stock prediction using Holt-Winters is shown in the following figure:
RMSE: 28.802996139471187
MAPE: 8.841081157580465

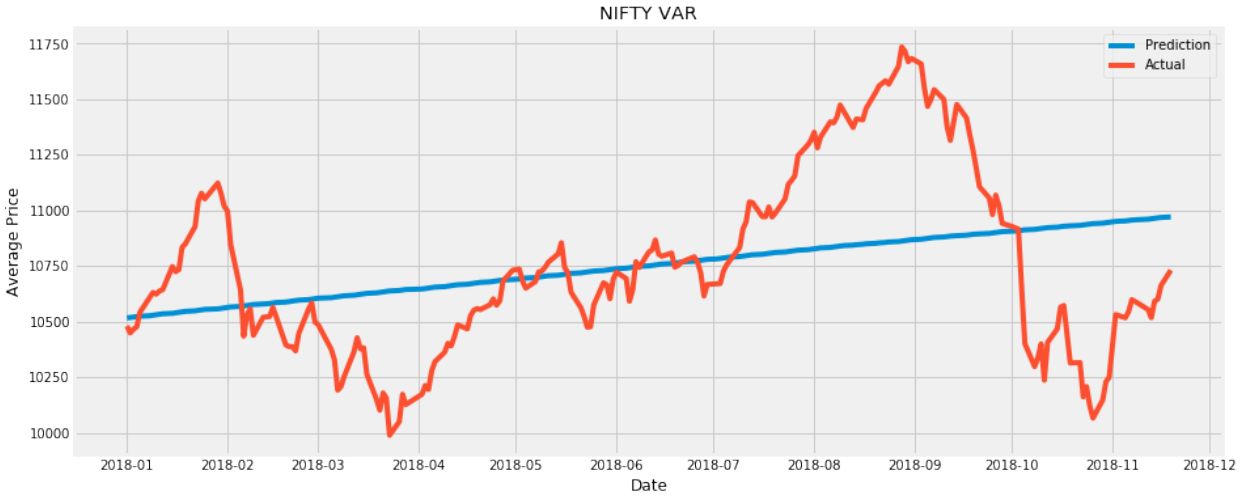


As can be seen from the RMSE and MAPE of the above models, they cannot be effectively used for predicting the stock data and has been only used as analytical tools to find out the underlying structure and seasonality of data. This is explained in the next section.

- **Features:**
  - In time-series prediction, the features of the data are the data points from previous time steps. The number of such time-steps is represented by the hyper-parameter 'look-back' which is the seasonality parameter in SARIMAX found while executing the pre-analysis section. In our case, the seasonality of the data was observed to be of five time steps. Hence, the feature set of the data for all the experimental models were $t, t_{-1}.t_{-2}, t_{-3}, t_{-4}$.
  - Since the experimental setup includes NIFTY as an exogenous variable, it has been used as another feature to train the experimental models.
- **Base Learners:**

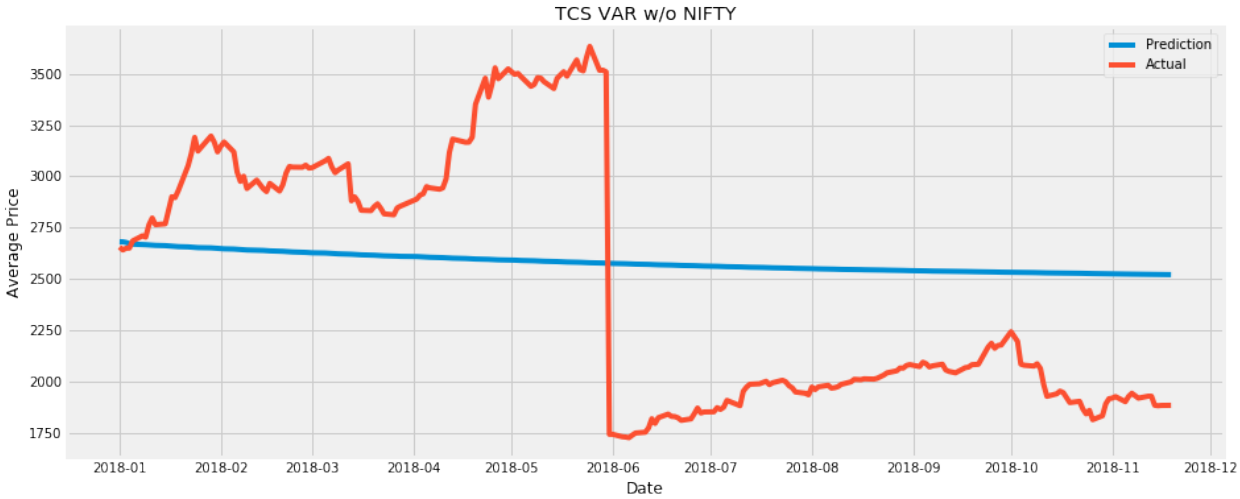
VAR (Vector Autoregression) model has been used as the base learner. This is because it is the simplest linear regressive statistical model available for multivariate time-series prediction that accepts exogenous variables. The results obtained from VAR are given below:

The **NIFTY** prediction using **VAR** is shown in the following figure:
RMSE: 380.223977393043
MAPE: 2.7421566289649615

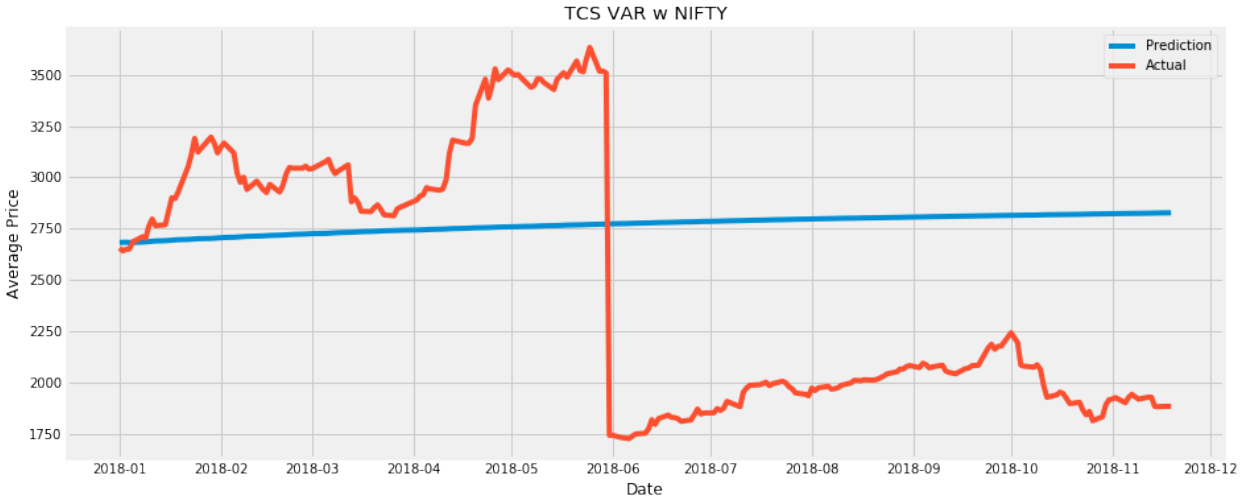


TCS stock prediction:

- Using **VAR without NIFTY** is shown in the following figure:
  RMSE: 585.6584321621915
  MAPE: 23.322750173938207

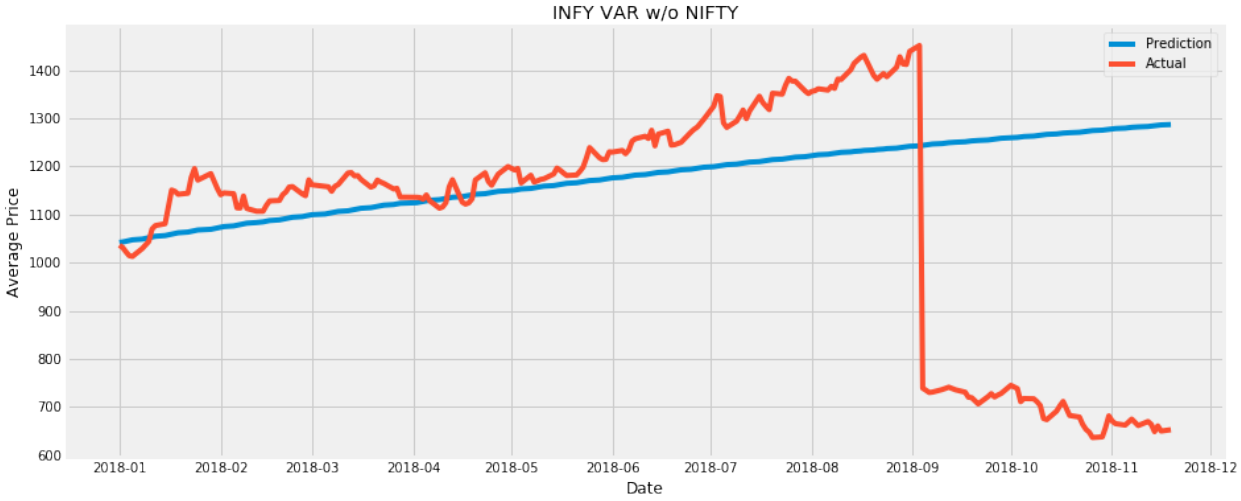


– Using **VAR with NIFTY** is shown in the following figure:
RMSE: 694.8295589437141
MAPE: 28.70074624609545

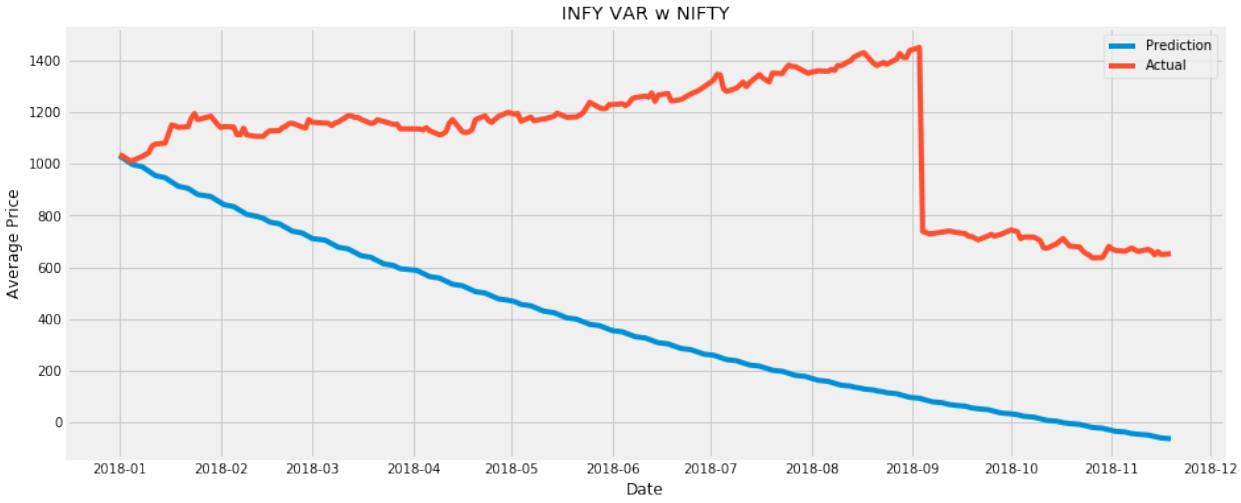


Infosys stock prediction:

– Using **VAR without NIFTY** is shown in the following figure:
RMSE: 284.7661268140489
MAPE: 23.260573235781695

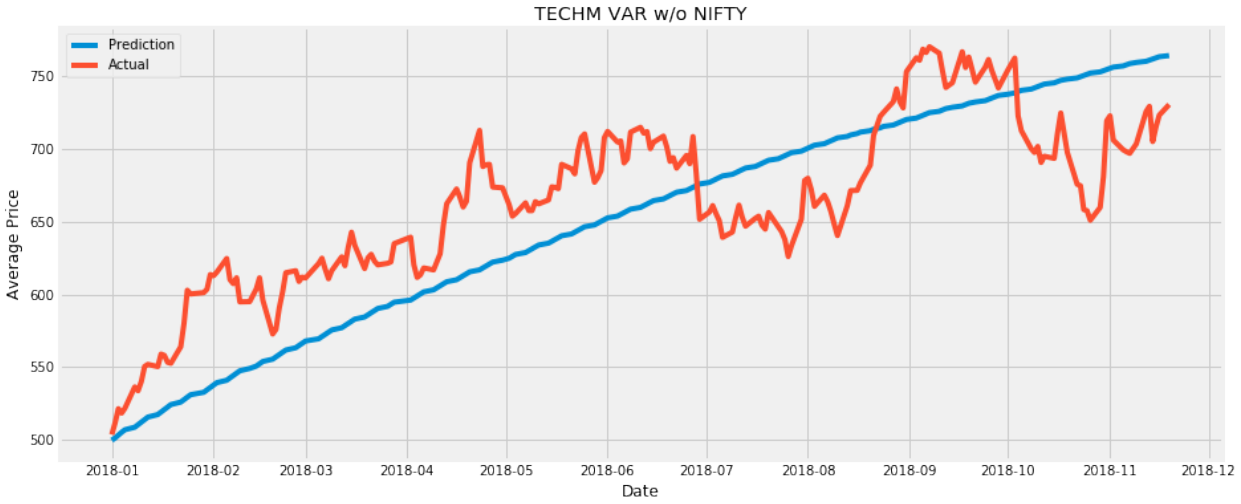


– Using **VAR with NIFTY** is shown in the following figure:
RMSE: 786.8385213075269
MAPE: 67.18116717624271

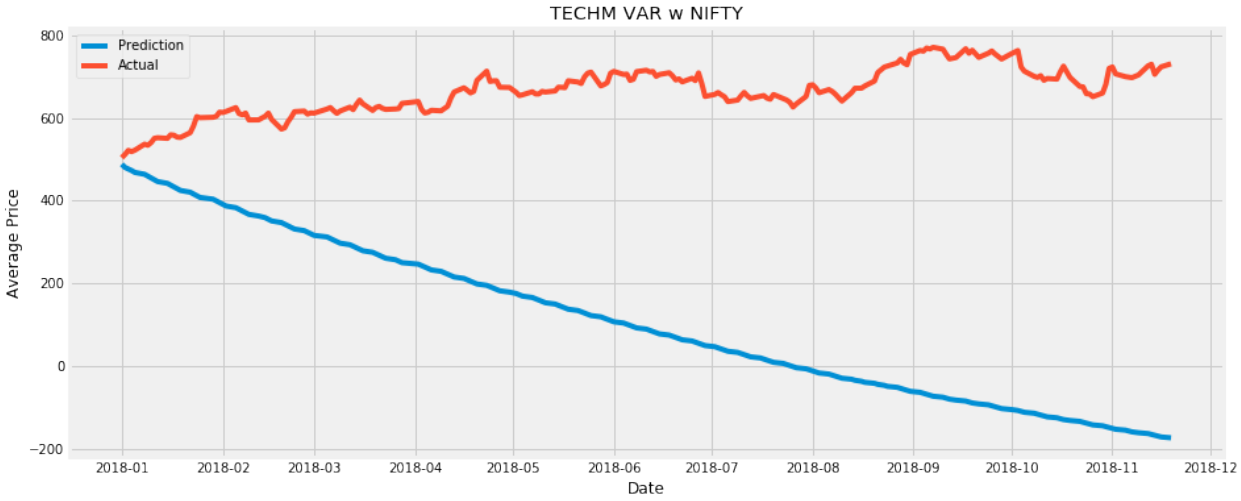


TECHM stock prediction:

– Using **VAR without NIFTY** is shown in the following figure:
RMSE: 44.744399434904814
MAPE: 6.1764025512634735

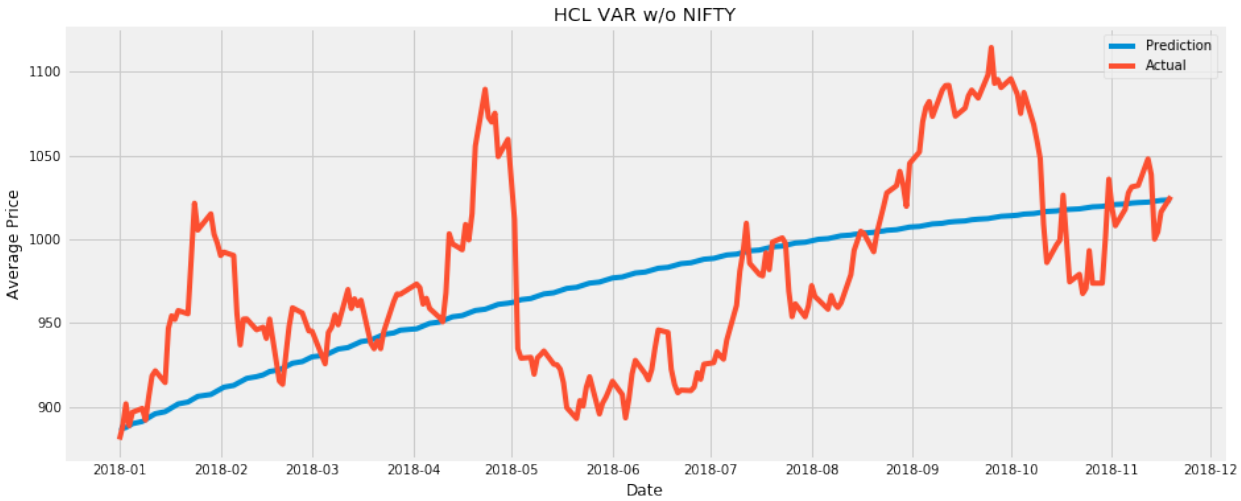


– Using **VAR with NIFTY** is shown in the following figure:
RMSE: 44.744399434904814
MAPE: 6.1764025512634735

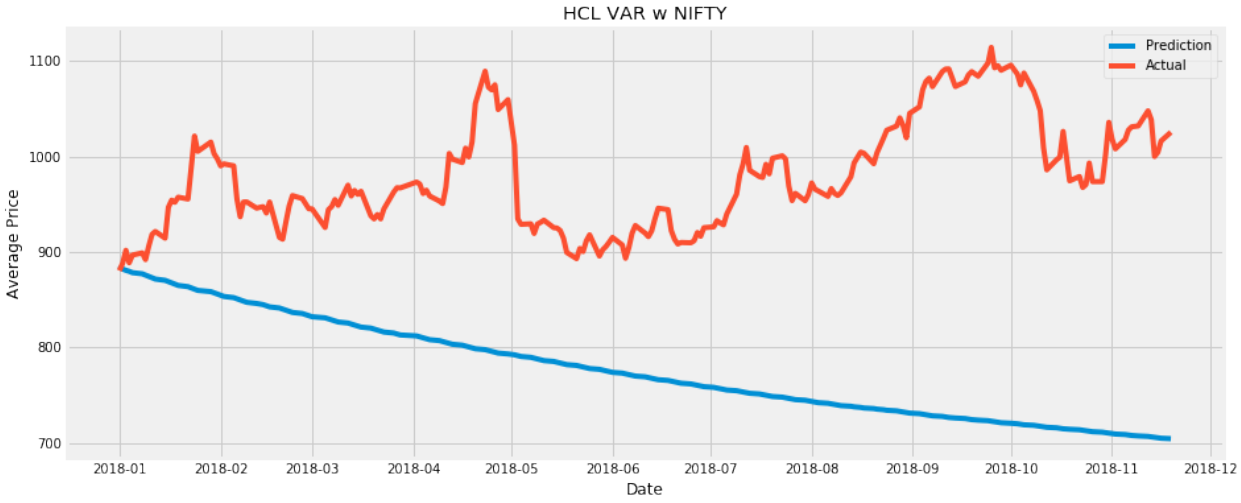


HCL stock prediction:

– Using **VAR without NIFTY** is shown in the following figure:
RMSE: 50.59836976690613
MAPE: 4.198930003821589

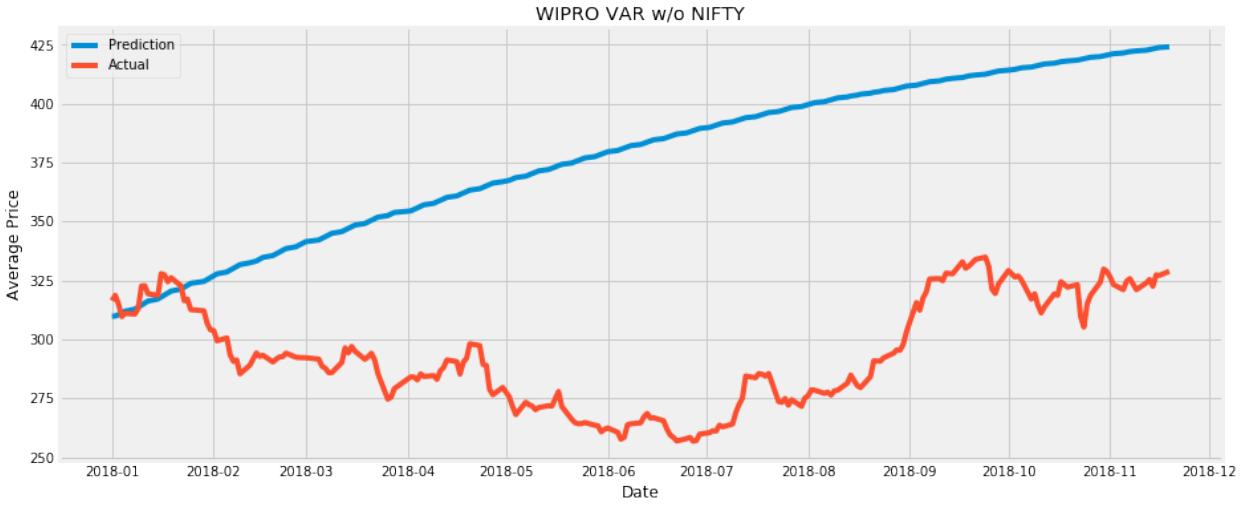


– Using **VAR with NIFTY** is shown in the following figure:
RMSE: 220.25381776460605
MAPE: 20.00959534000949

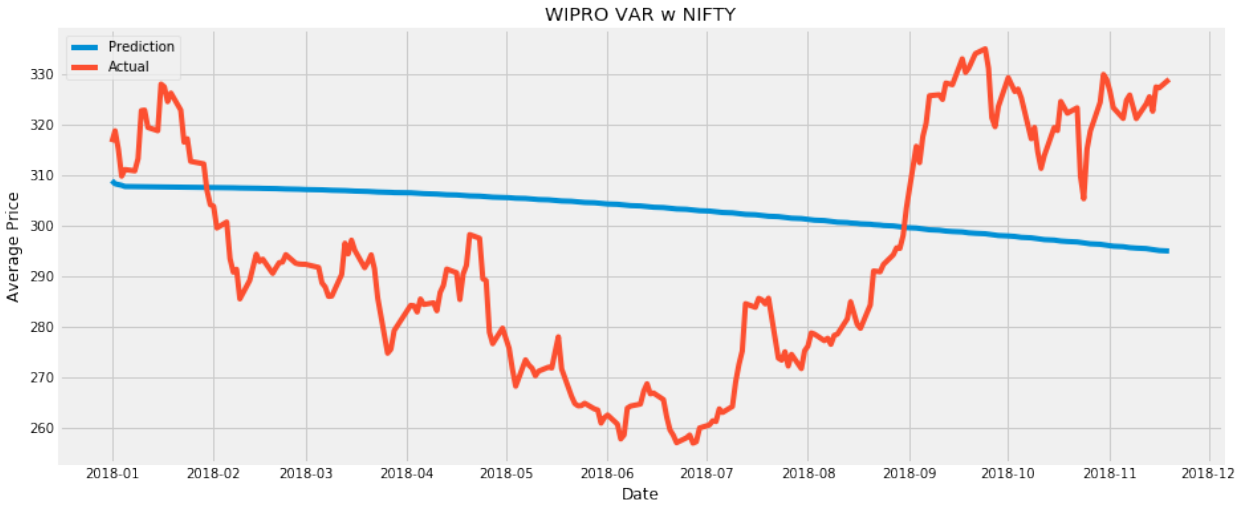


WIPRO stock prediction:

– Using **VAR without NIFTY** is shown in the following figure:
RMSE: 91.99122292936308
MAPE: 29.432541168796085



– Using **VAR with NIFTY** is shown in the following figure:
RMSE: 26.29028955525635
MAPE: 8.282934307028086



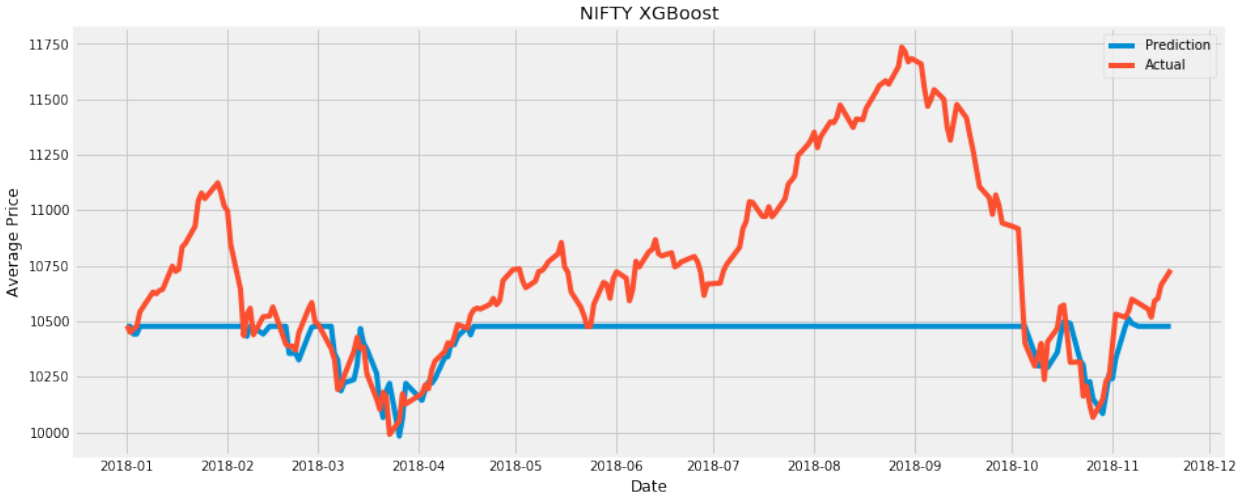
- **Experimental Models:**

  The experimental model set comprises of two Deep learning models and one machine learning model which is XGBoost. XGBoost has been known to give very good performance for time-series data, however, it has its limitations as it cannot predict the 'highs' very well. It predicts the highs as the highest value that it has seen during the training and as the testing data is from 2018 and training before that, it means that XGBoost cannot predict if the companies make a new all time high which can be seen in the results. The model also has multiple hyper-parameters that were tuned: number of estimators, maximum depth of each estimator (tree size) and learning rate.

  Similarly in the deep learning models we see that the disadvantage in the XGBoost algorithm is overcome and these models fair extremely well even in the case of predicting all-time-highs. The hyper-parameter that was tuned is batch size which takes into account the number data points that are used in each batch. Other hyper-parameters such as number of epochs and dropout probability were left un-tuned due to various reasons, such as, existing model's already good accuracy and time constraints. Multiple networks were tried and the best was selected for implementation.

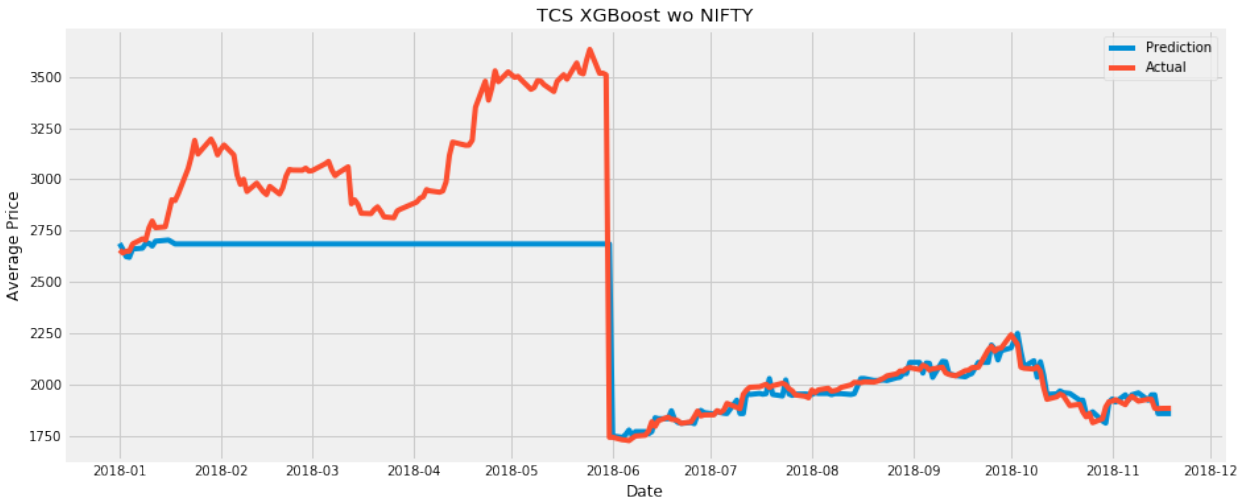
  It has to be noted that both of these models predict the drastic, overnight fall in the value of the stock due to an announcement of a share-split. A share-split is where the company decides to split the value of the share, say 1:2, so every investor who own one share will have 2 at half the previous value. This prediction was made by all the three models almost to the day, which is very impressive given that they have only one feature to work with. The prediction plots of all three are as follows:

– XGBOOST:
The **NIFTY** prediction using **XGBOOST** is shown in the following figure:
RMSE: 380.223977393043
MAPE: 2.7421566289649615

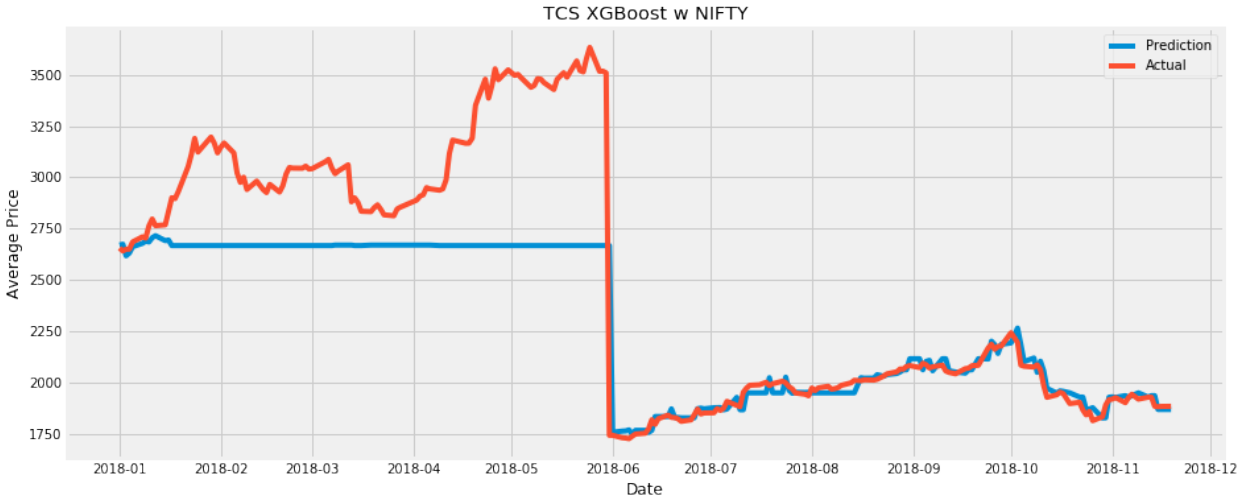


TCS stock prediction:

* Using **XGBOOST without NIFTY** is shown in the following figure:
RMSE: 585.6584321621915
MAPE: 23.322750173938207

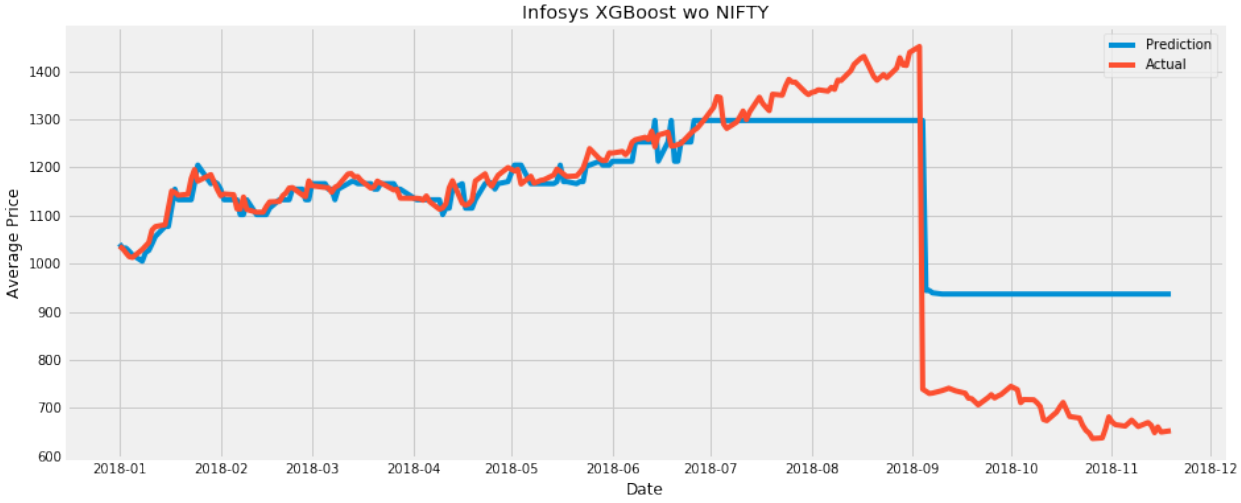


* Using **XGBOOST with NIFTY** is shown in the following figure:
  RMSE: 356.7005067213686
  MAPE: 7.196866770811115

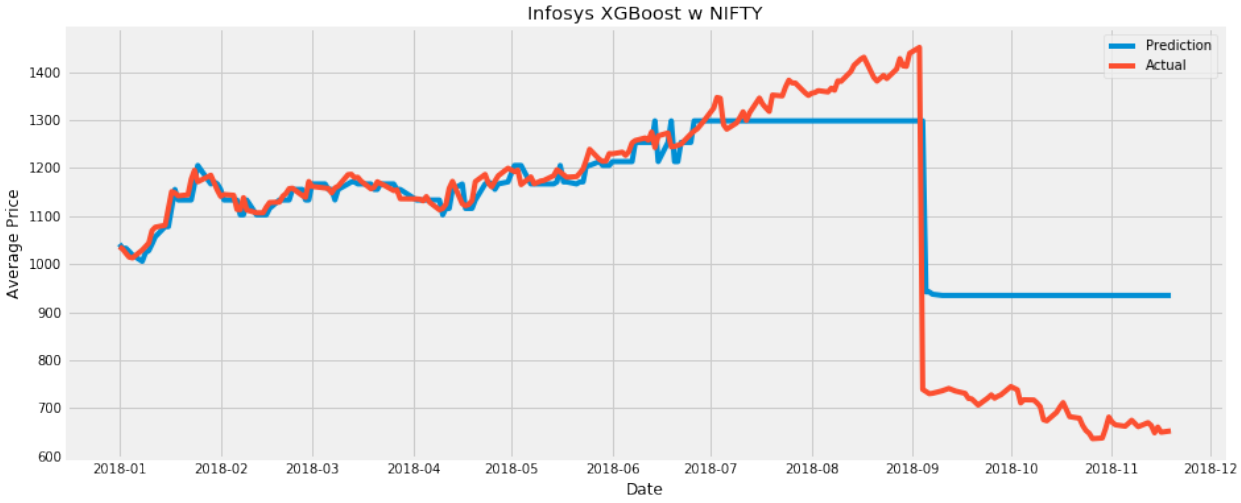


Infosys stock prediction:

* Using **XGBOOST without NIFTY** is shown in the following figure:
  RMSE: 128.03554614123877
  MAPE: 10.033083227101745

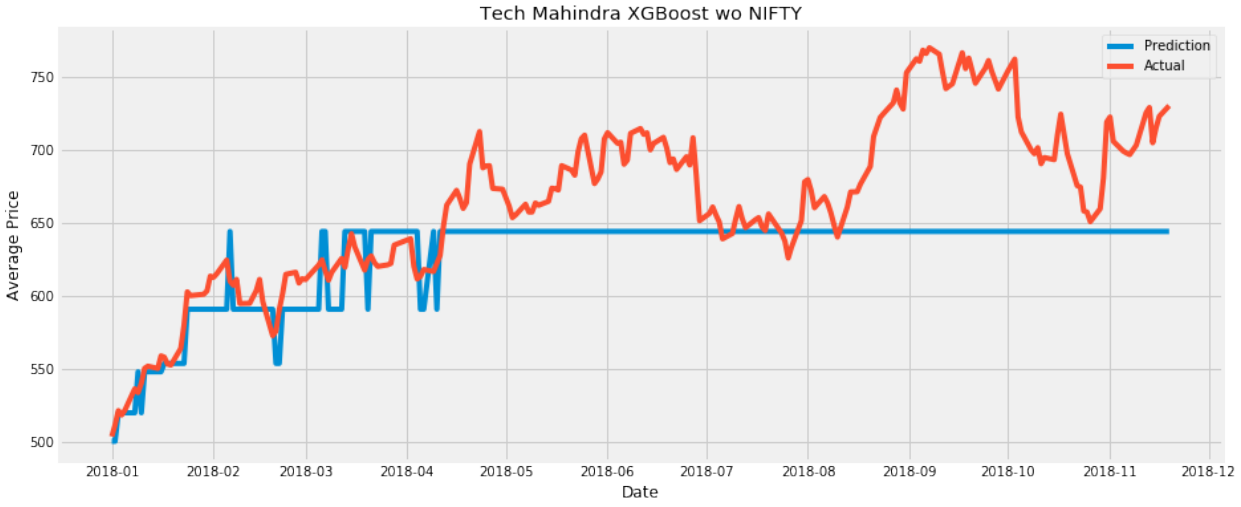


* Using **XGBOOST with NIFTY** is shown in the following figure:
  RMSE: 127.09807409895662
  MAPE: 9.946749769781862

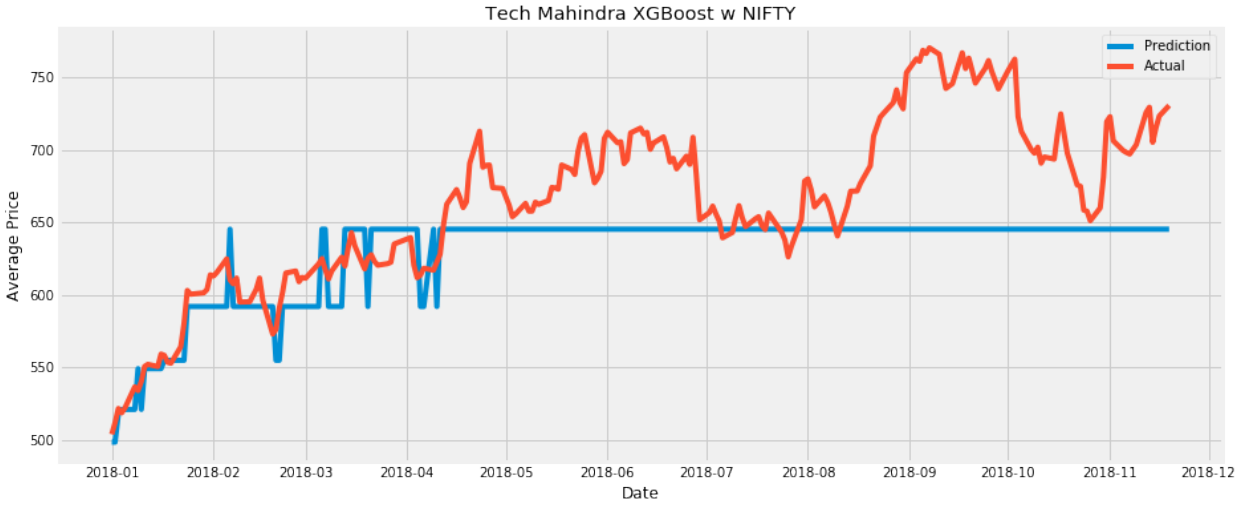


TECHM stock prediction:

* Using **XGBOOST without NIFTY** is shown in the following figure:
  RMSE: 51.476791532907335
  MAPE: 5.630836708111272

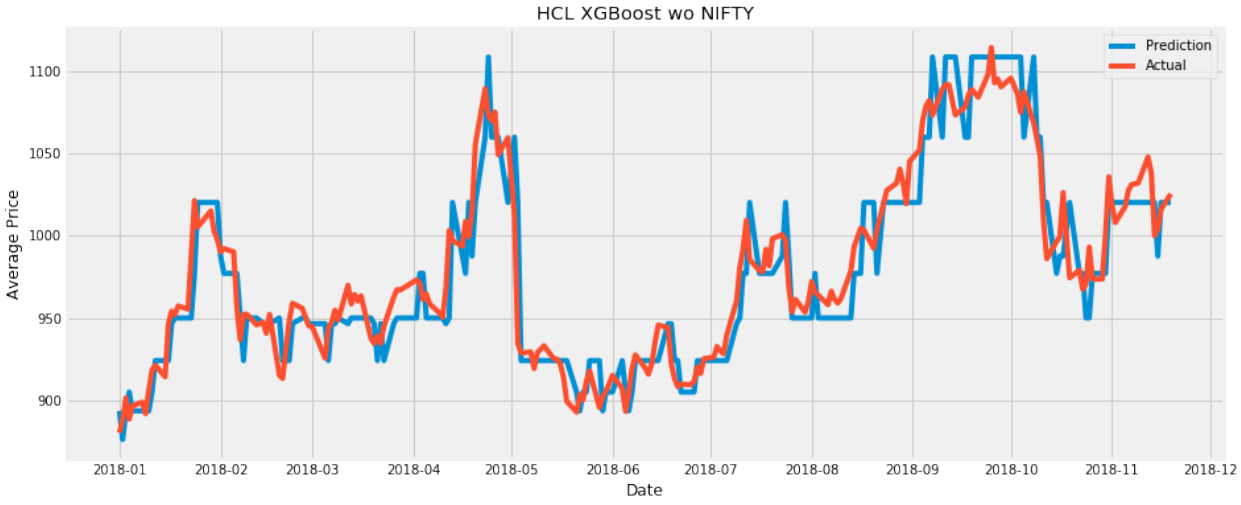


* Using **XGBOOST with NIFTY** is shown in the following figure:
  RMSE: 50.89305467117587
  MAPE: 5.547783903590692

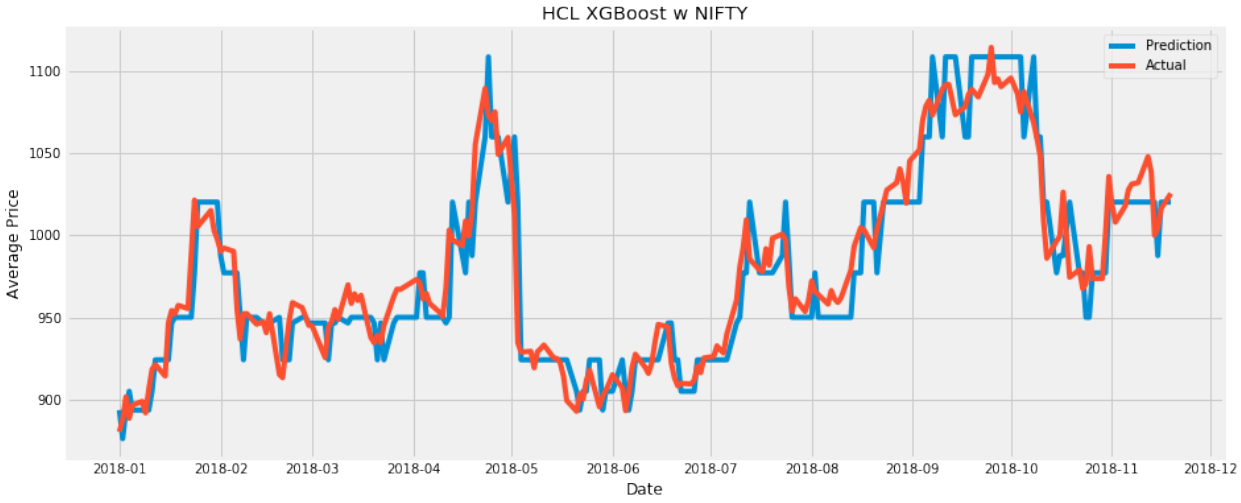


HCL stock prediction:

* Using **XGBOOST without NIFTY** is shown in the following figure:
  RMSE: 18.897340318520328
  MAPE: 1.4897004475757951

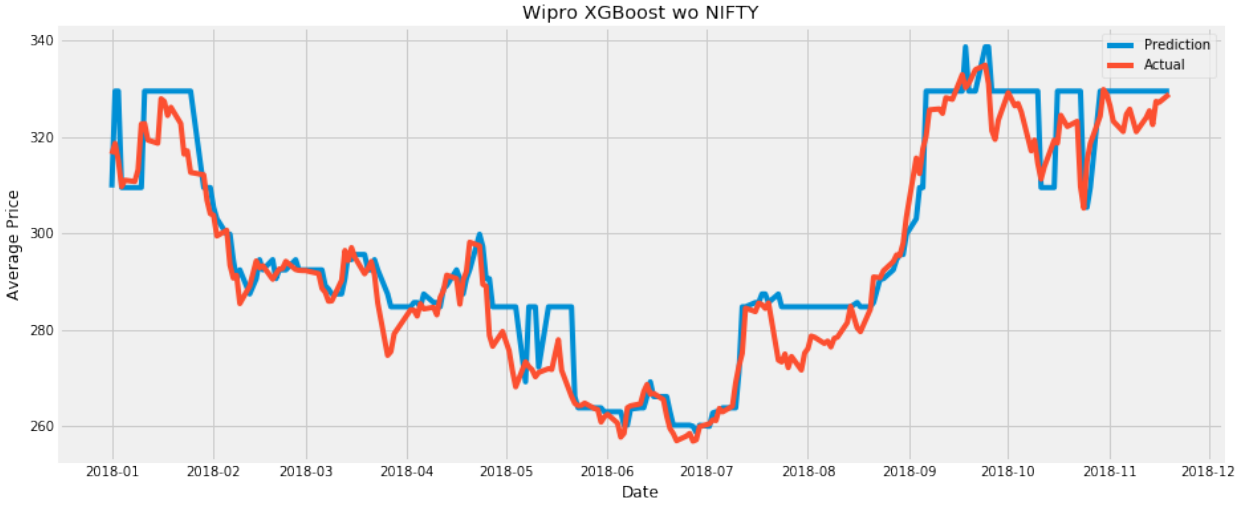


* Using **XGBOOST with NIFTY** is shown in the following figure:
  RMSE: 18.897340318520328
  MAPE: 1.4897004475757951

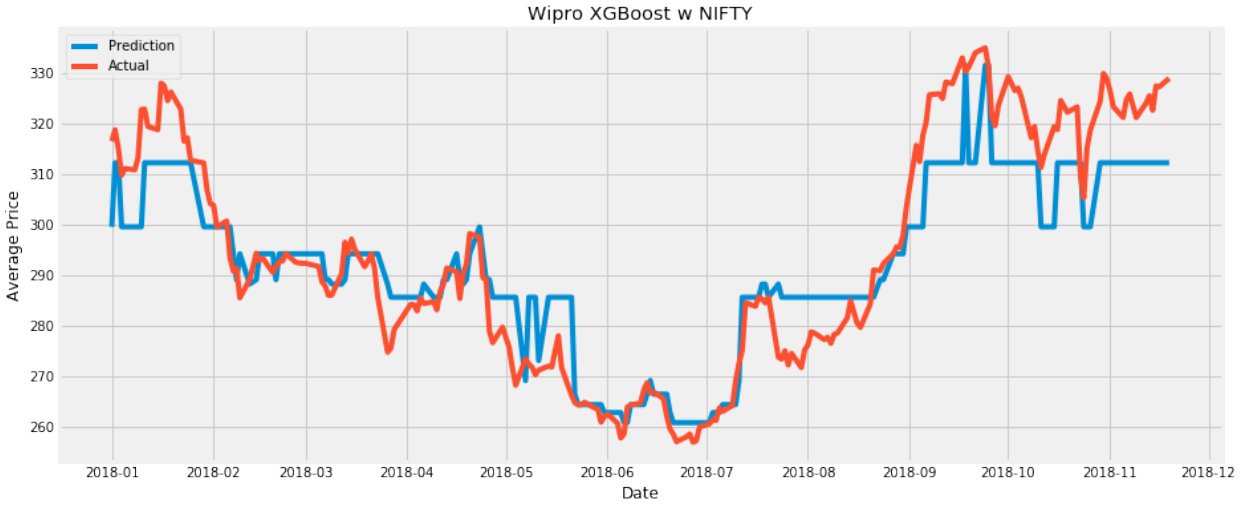


WIPRO stock prediction:

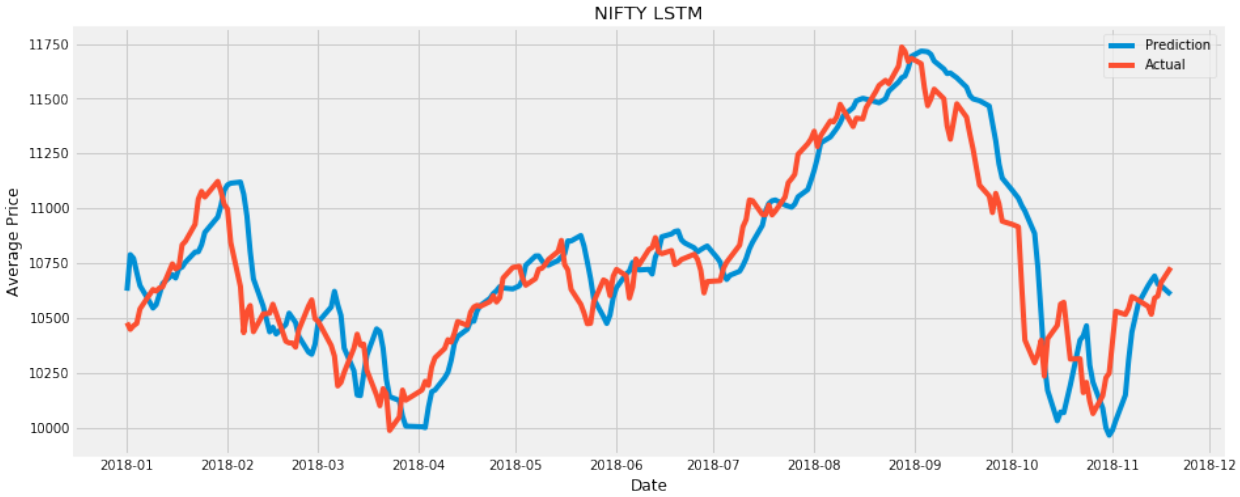
* Using **XGBOOST without NIFTY** is shown in the following figure:
  RMSE: 6.434699267260752
  MAPE: 1.62607597466029



* Using **XGBOOST with NIFTY** is shown in the following figure:
  RMSE: 8.908333537388739
  MAPE: 2.260448354911617

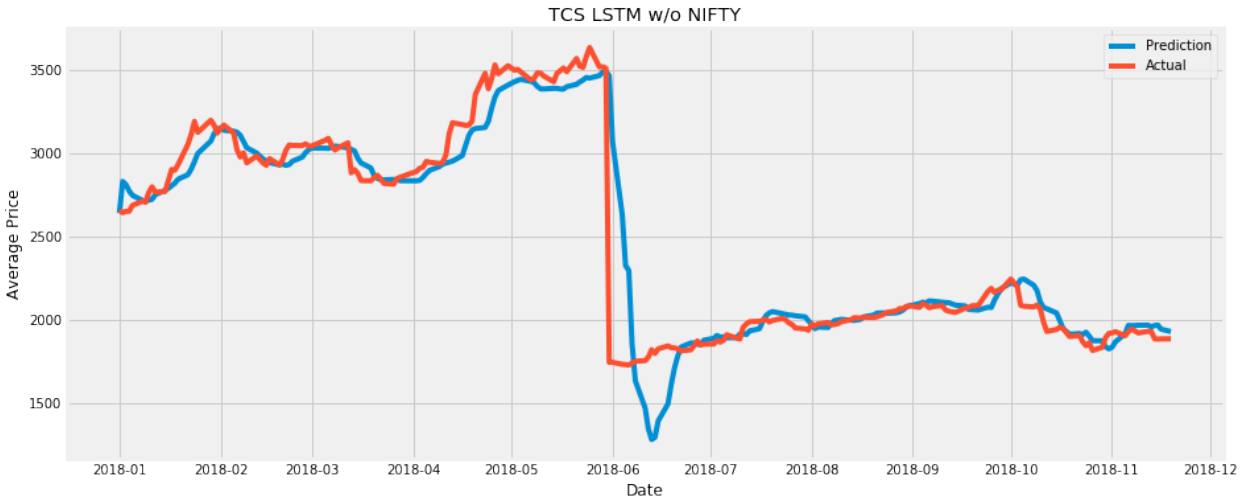


– LSTM:
  The **NIFTY** prediction using **LSTM** is shown in the following figure:
  RMSE: 189.35024928093634
  MAPE: 1.362965300338482

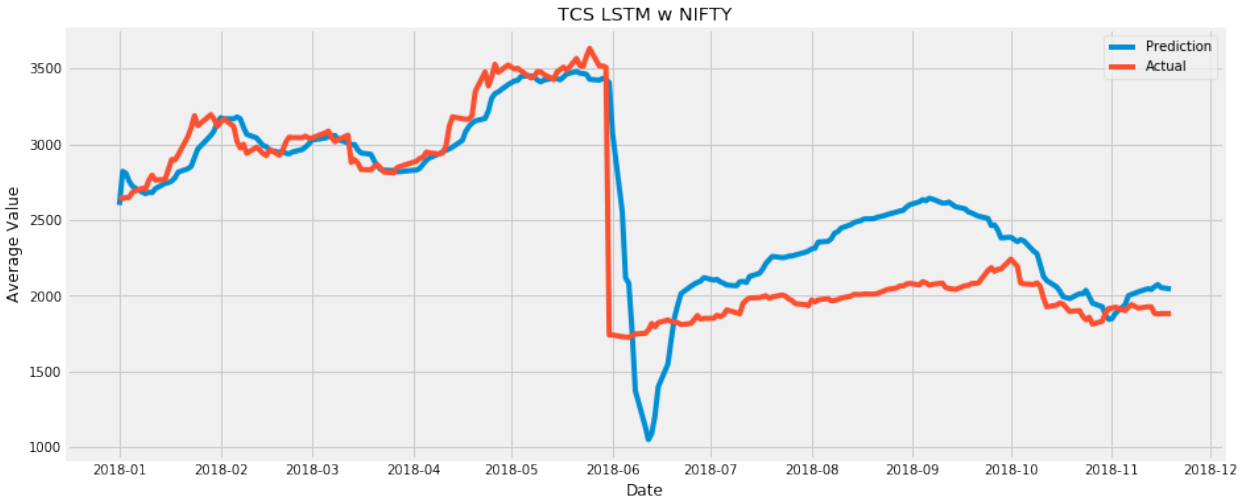


TCS stock prediction:

* Using **LSTM without NIFTY** is shown in the following figure:
  RMSE: 200.18207572026213
  MAPE: 4.224200539050362

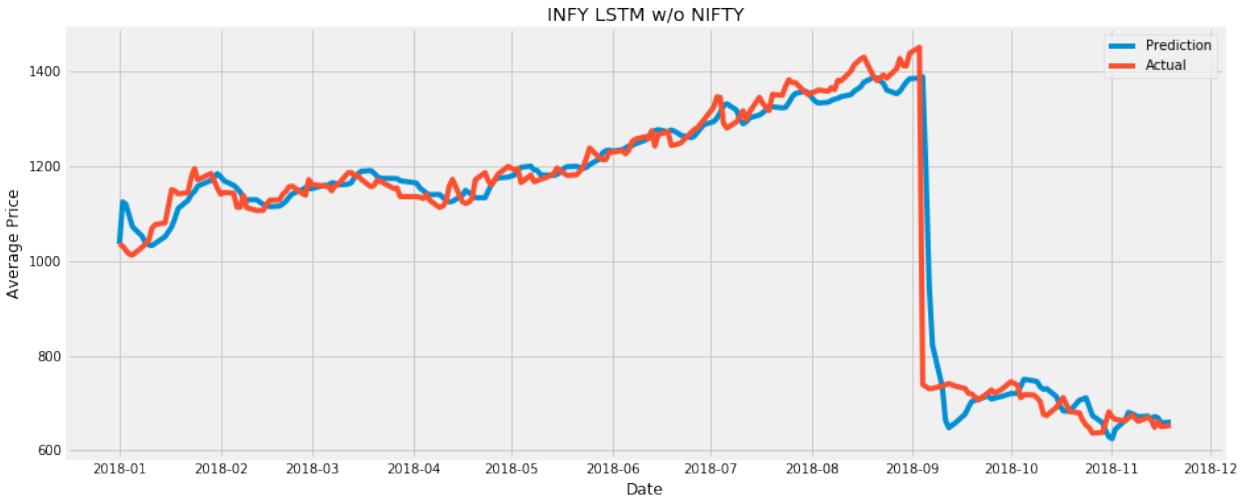


* Using **LSTM with NIFTY** is shown in the following figure:
  RMSE: 300.2988351463691
  MAPE: 10.030874490620329

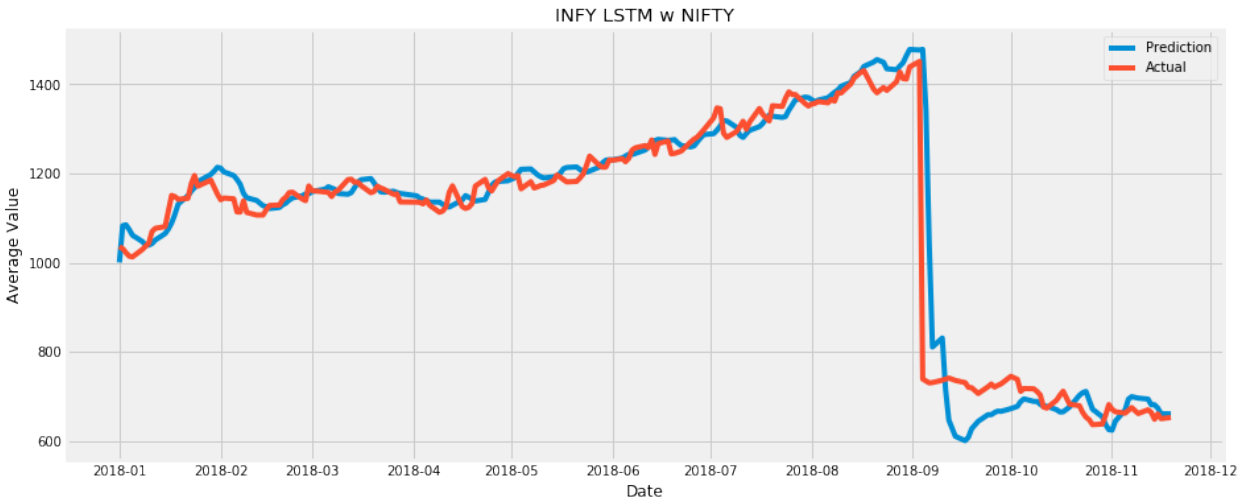


Infosys stock prediction:

* Using **LSTM without NIFTY** is shown in the following figure:
  RMSE: 63.68811883163279
  MAPE: 3.122195702938981

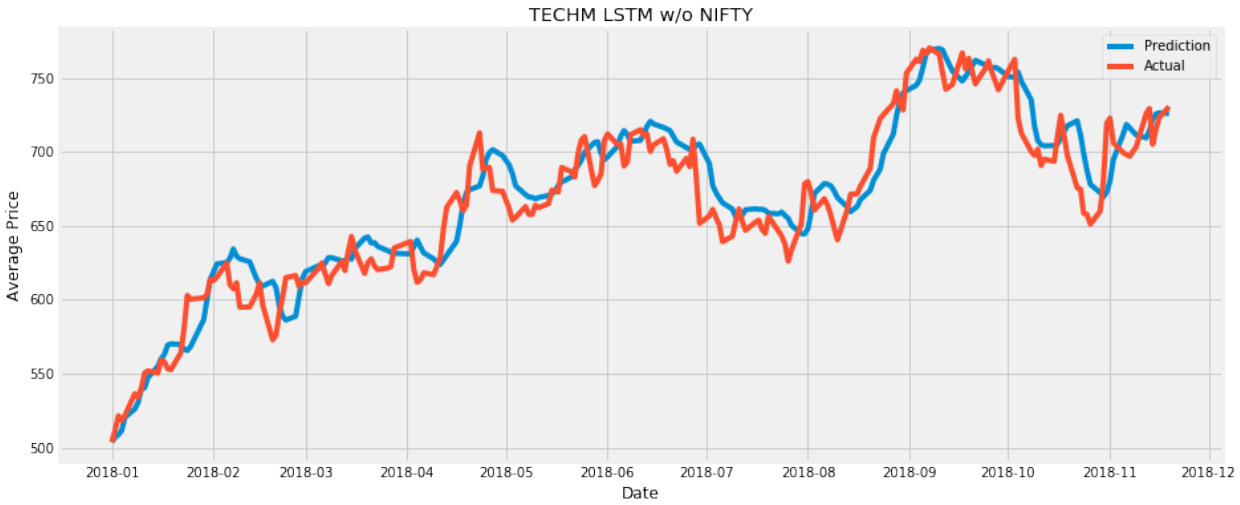


* Using **LSTM with NIFTY** is shown in the following figure:
  RMSE: 76.46804634955403
  MAPE: 3.7166491663927603

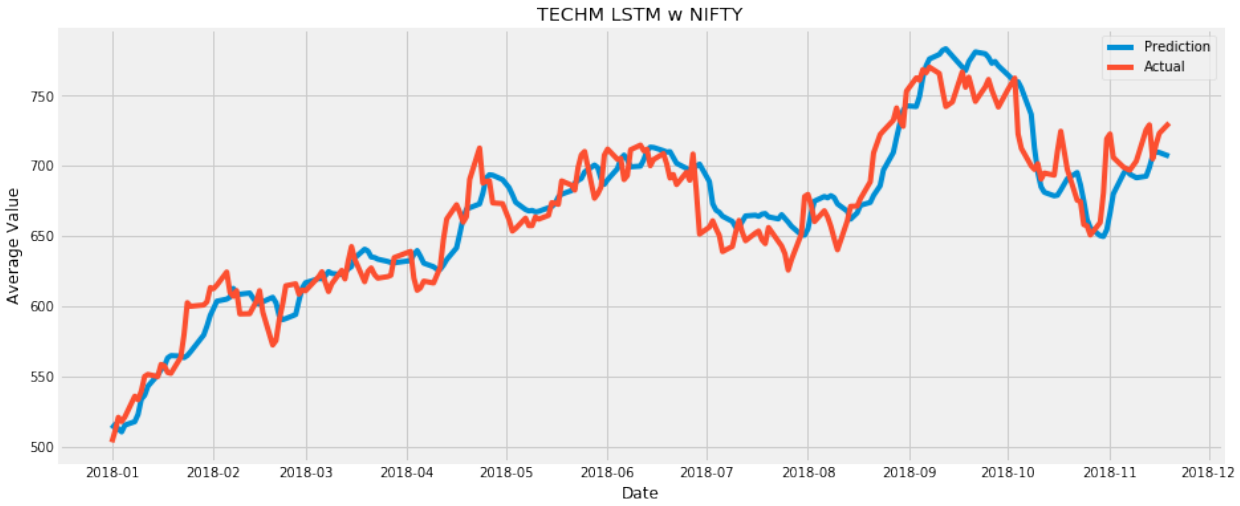


TECHM stock prediction:

* Using **LSTM without NIFTY** is shown in the following figure:
  RMSE: 17.921052036645733
  MAPE: 2.181580089802378

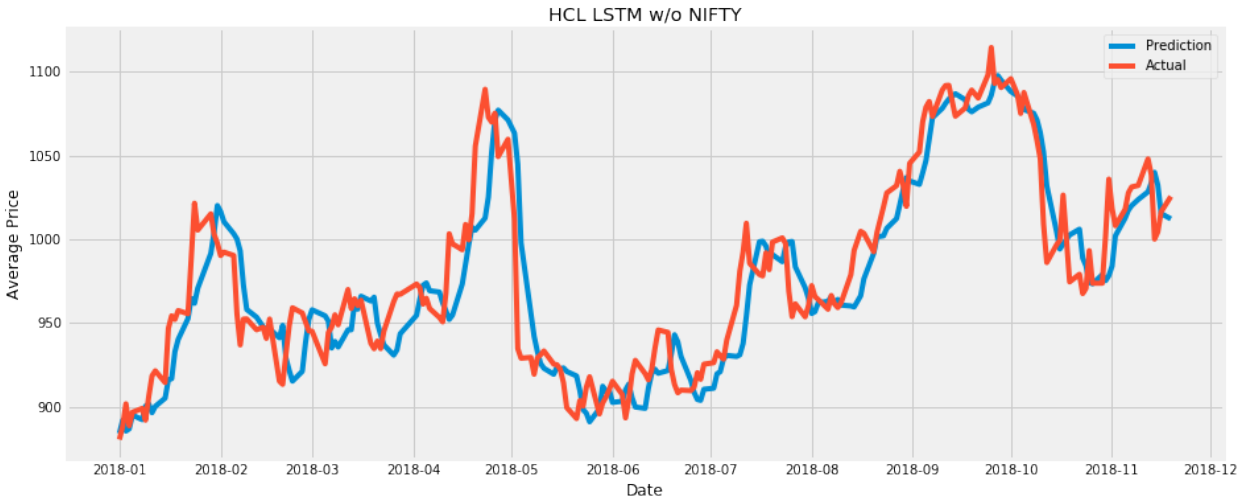


* Using **LSTM with NIFTY** is shown in the following figure:
  RMSE: 18.371912590164854
  MAPE: 2.182116759368078

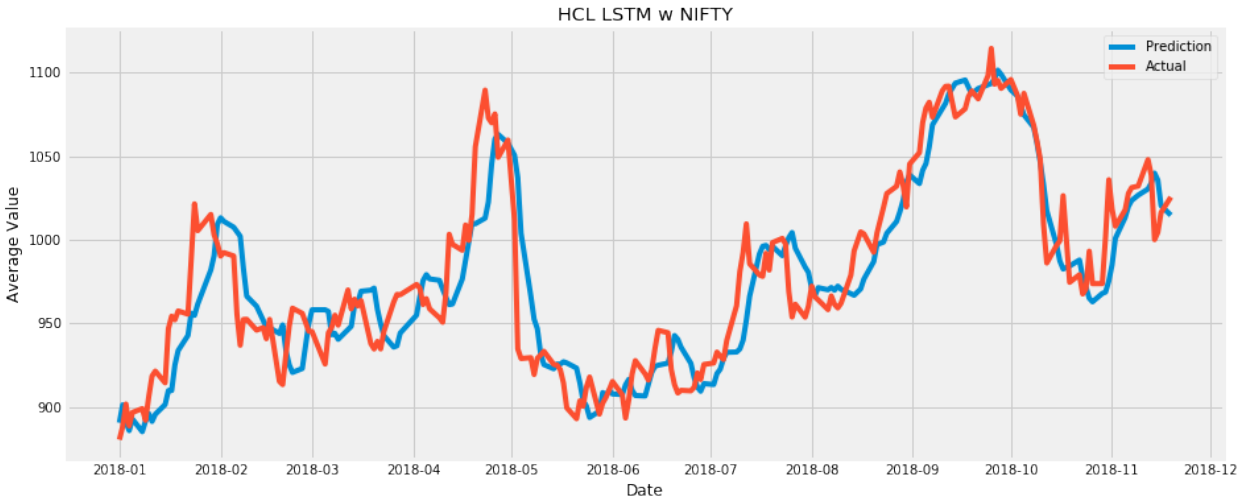


HCL stock prediction:

* Using **LSTM without NIFTY** is shown in the following figure:
  RMSE: 23.571958318134538
  MAPE: 1.7867616855348731

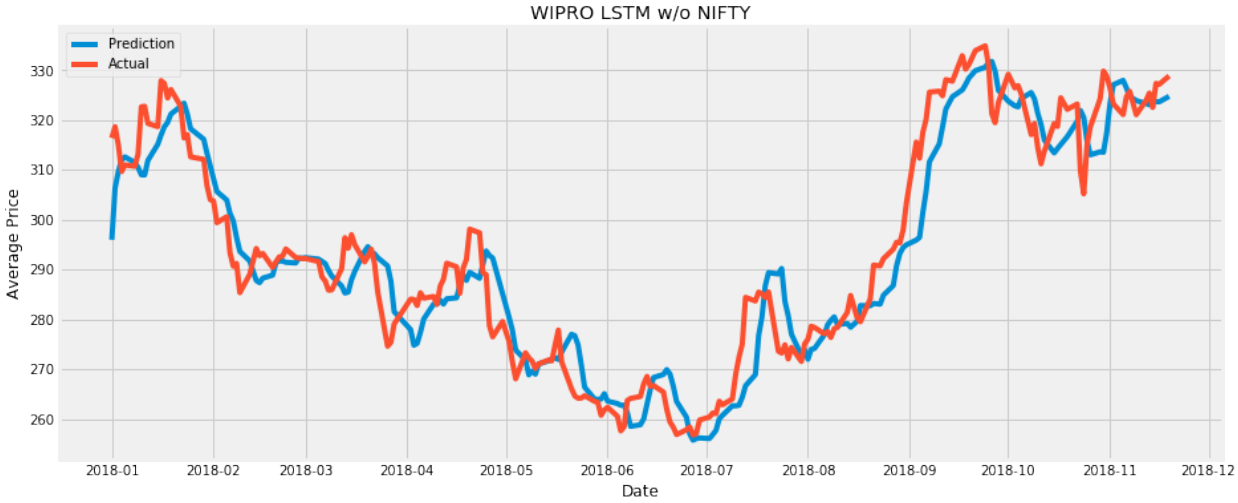


* Using **LSTM with NIFTY** is shown in the following figure:
  RMSE: 23.94276702370191
  MAPE: 1.8279734591294297

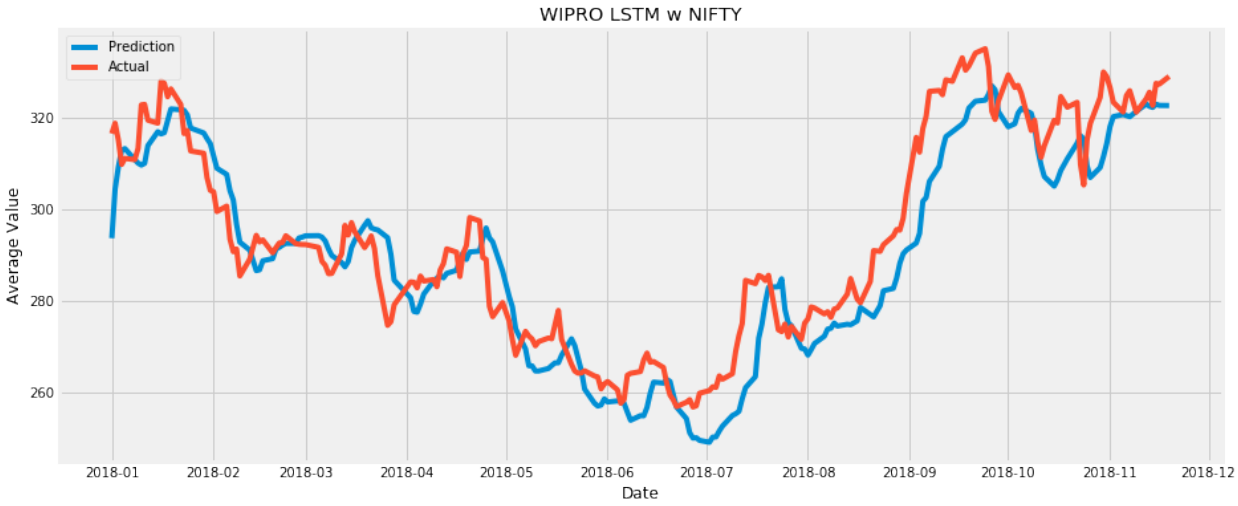


WIPRO stock prediction:

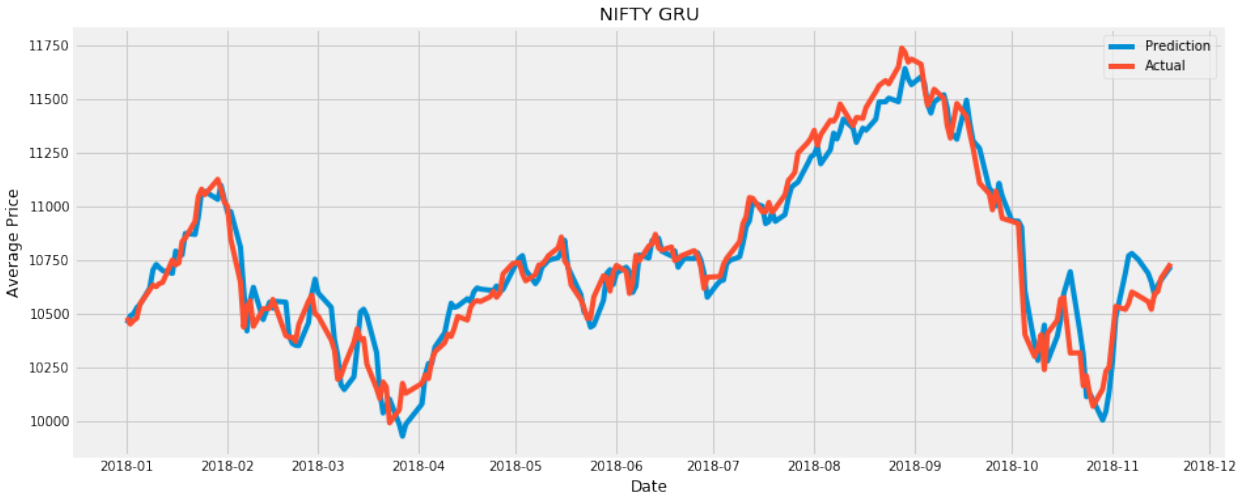
* Using **LSTM without NIFTY** is shown in the following figure:
  RMSE: 6.910655929197406
  MAPE: 1.8365334578121748



* Using **LSTM with NIFTY** is shown in the following figure:
  RMSE: 8.597425201561196
  MAPE: 2.3854286757055037

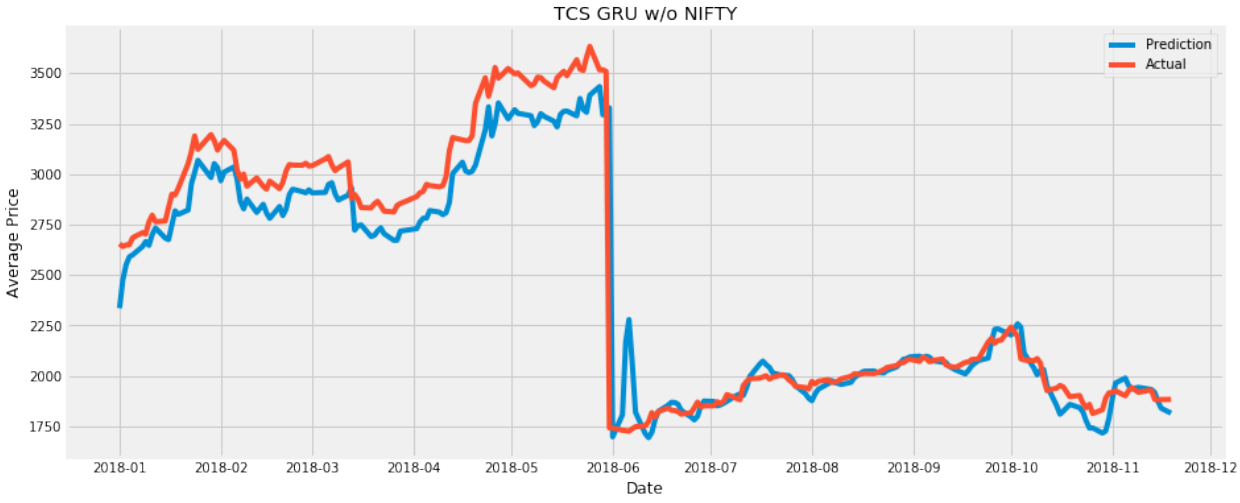


– GRU:
The **NIFTY** prediction using **GRU** is shown in the following figure:
RMSE: 93.81802940008629
MAPE: 0.6913757689994254

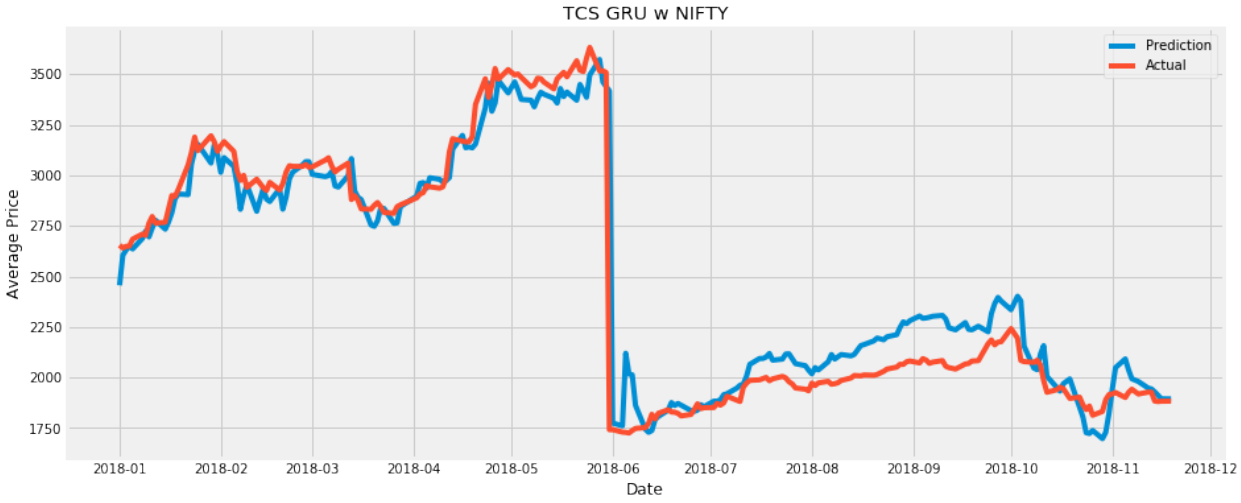


TCS stock prediction:

* Using **GRU without NIFTY** is shown in the following figure:
RMSE: 168.26441100954486
MAPE: 4.079932077536251

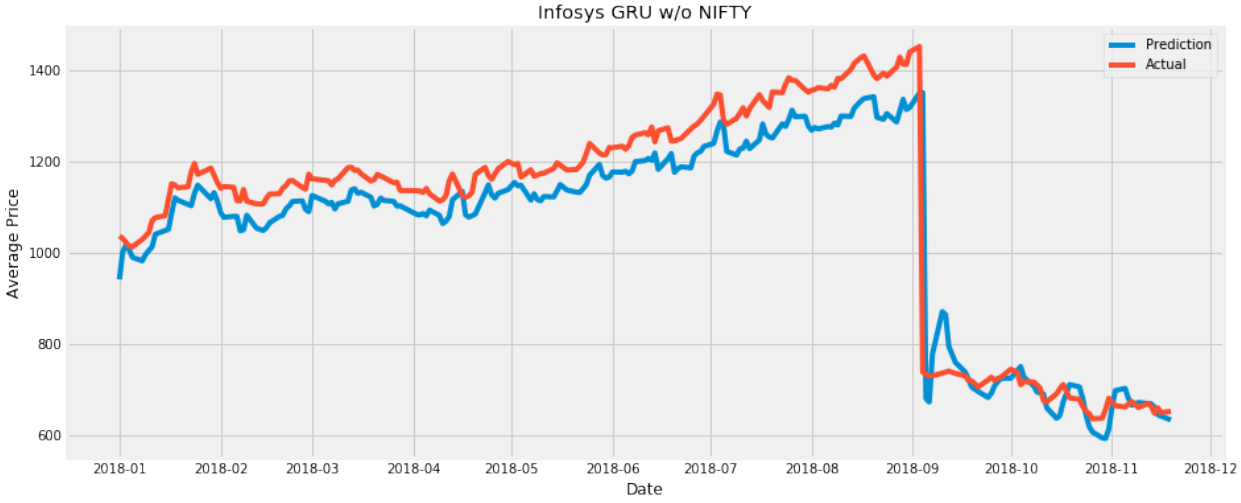


* Using **GRU with NIFTY** is shown in the following figure:
  RMSE: 159.97215954810648
  MAPE: 4.328941096073781

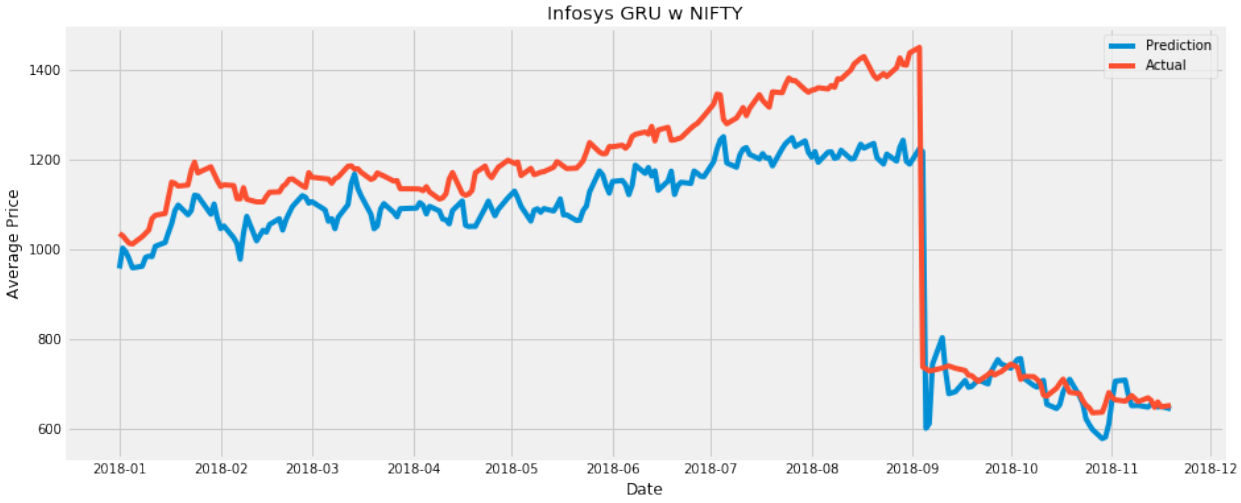


Infosys stock prediction:

* Using **GRU without NIFTY** is shown in the following figure:
  RMSE: 72.07669430453187
  MAPE: 5.005412036394907

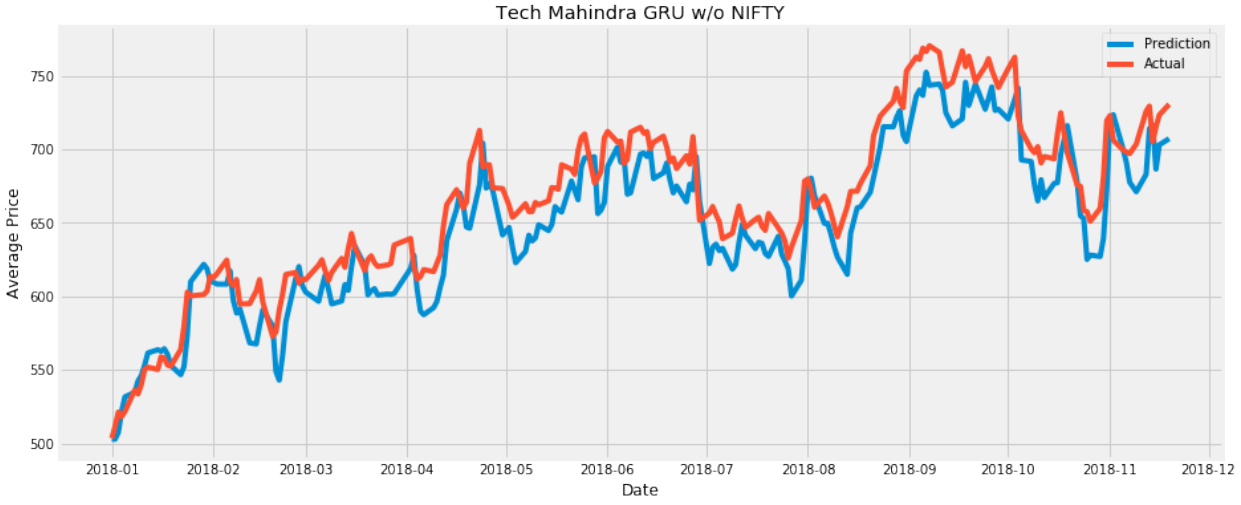


* Using **GRU with NIFTY** is shown in the following figure:
  RMSE: 101.09426921023862
  MAPE: 7.231762266928447

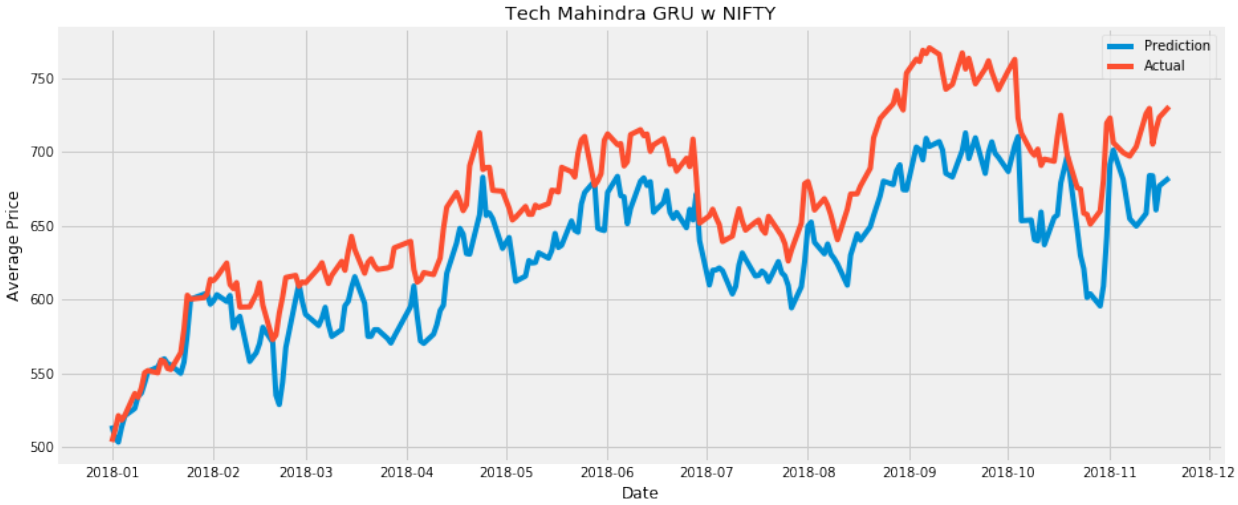


TECHM stock prediction:

* Using **GRU without NIFTY** is shown in the following figure:
  RMSE: 22.068647046392297
  MAPE: 2.833410264689146

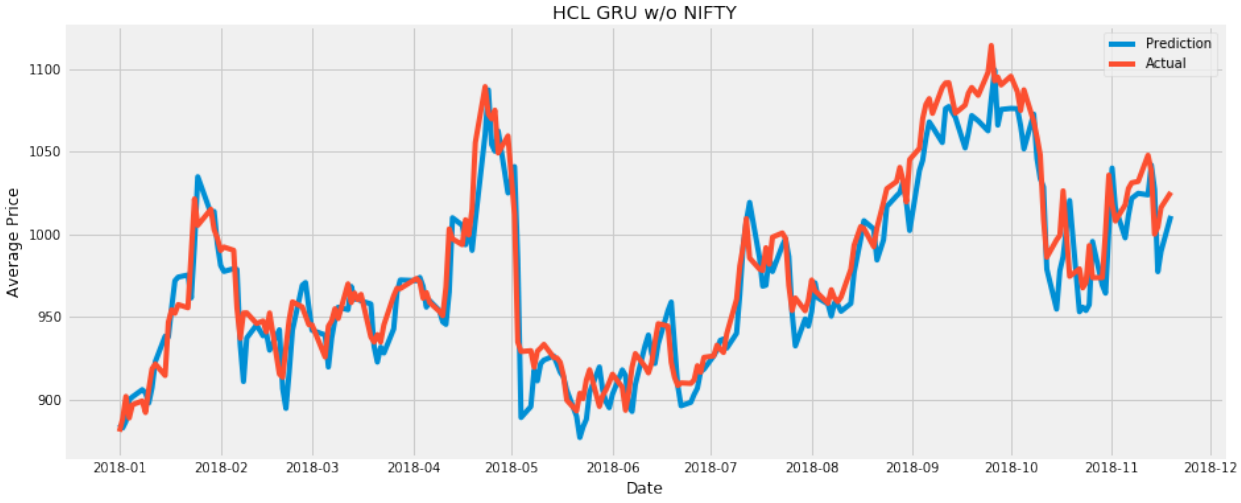


* Using **GRU with NIFTY** is shown in the following figure:
  RMSE: 39.60142960542171
  MAPE: 5.181458330231272

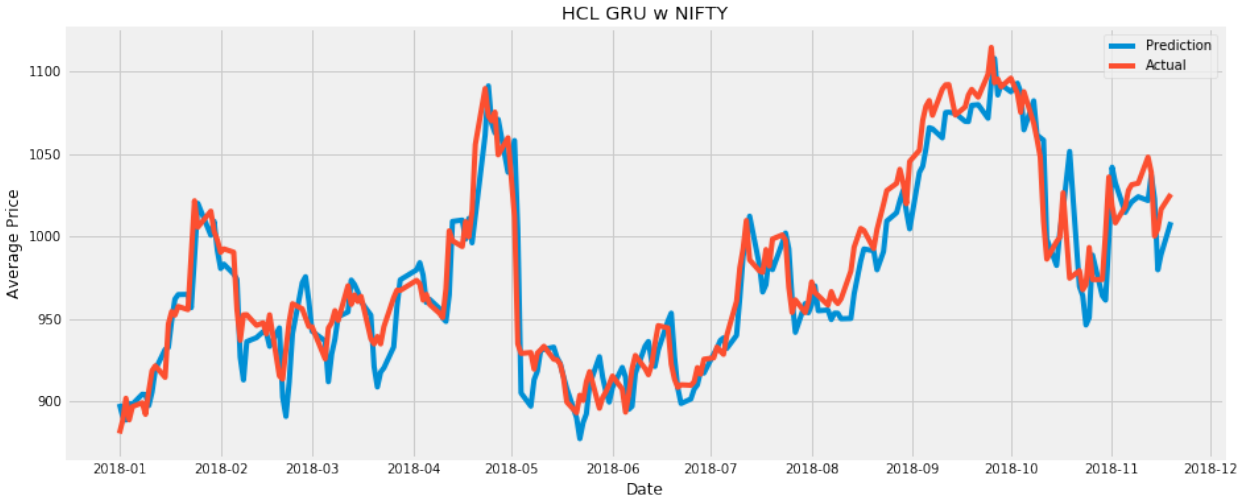


HCL stock prediction:

* Using **GRU without NIFTY** is shown in the following figure:
  RMSE: 18.55138741470192
  MAPE: 1.50993088856868

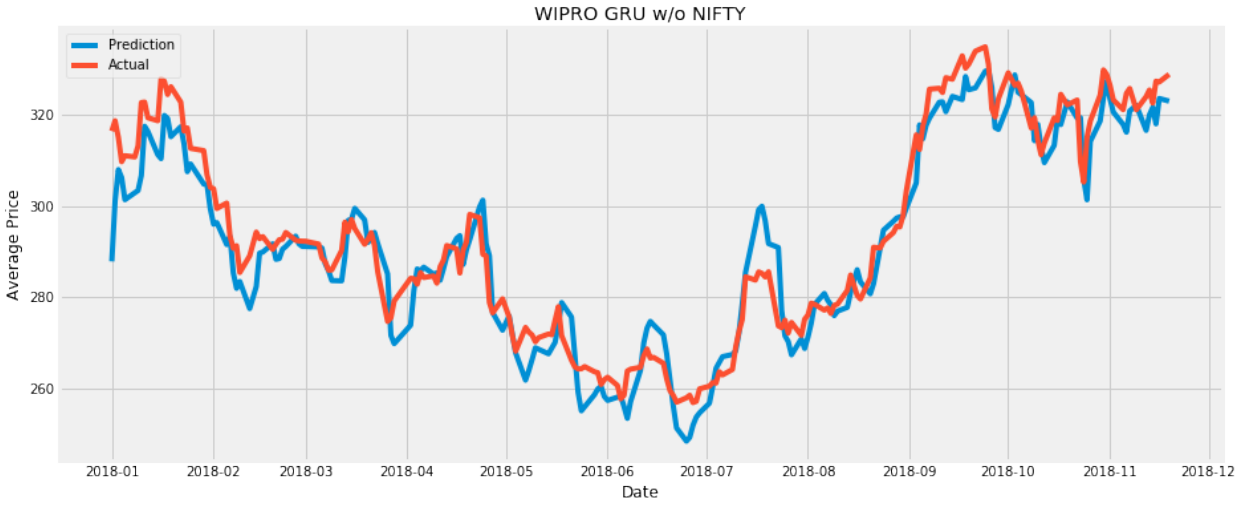


* Using **GRU with NIFTY** is shown in the following figure:
  RMSE: 18.916583864554838
  MAPE: 1.5110284816595936

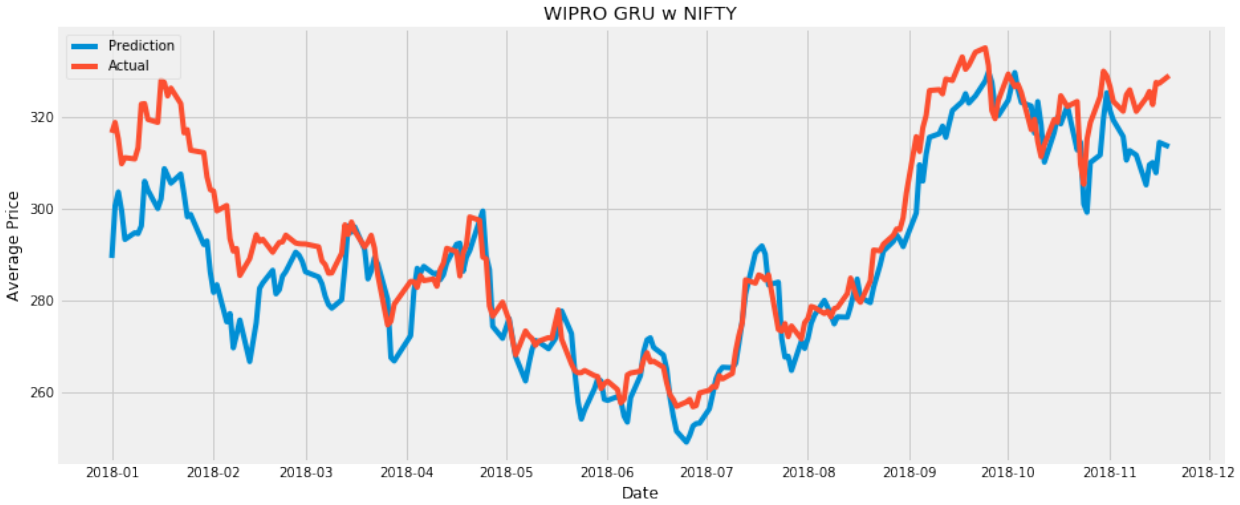


WIPRO stock prediction:

* Using **GRU without NIFTY** is shown in the following figure:
  RMSE: 6.266118930719043
  MAPE: 1.648842915024443



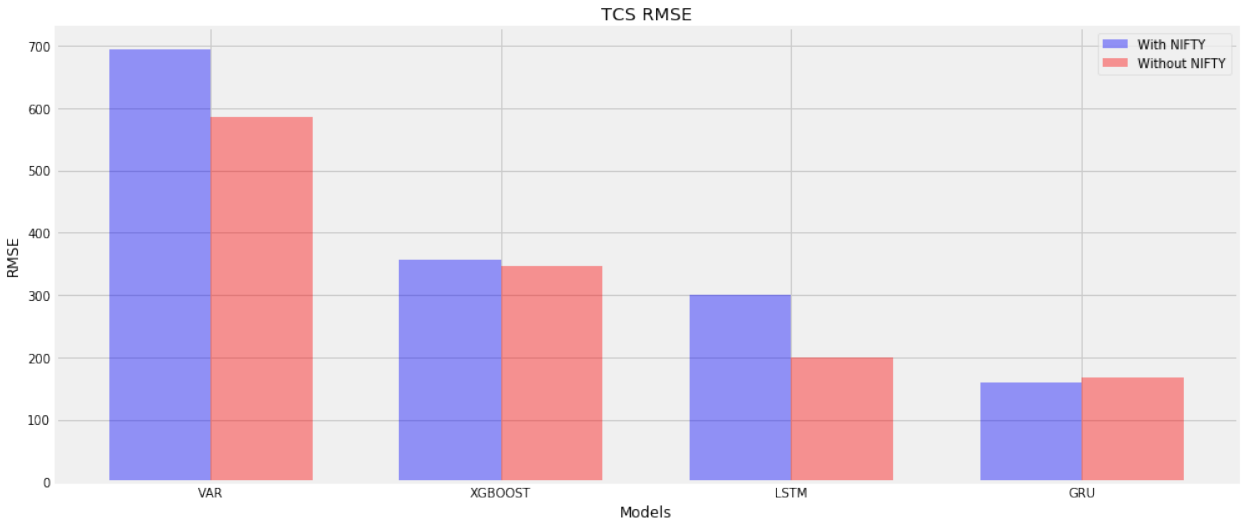
* Using **GRU with NIFTY** is shown in the following figure:
  RMSE: 9.376637385138418
  MAPE: 2.3760725854740974



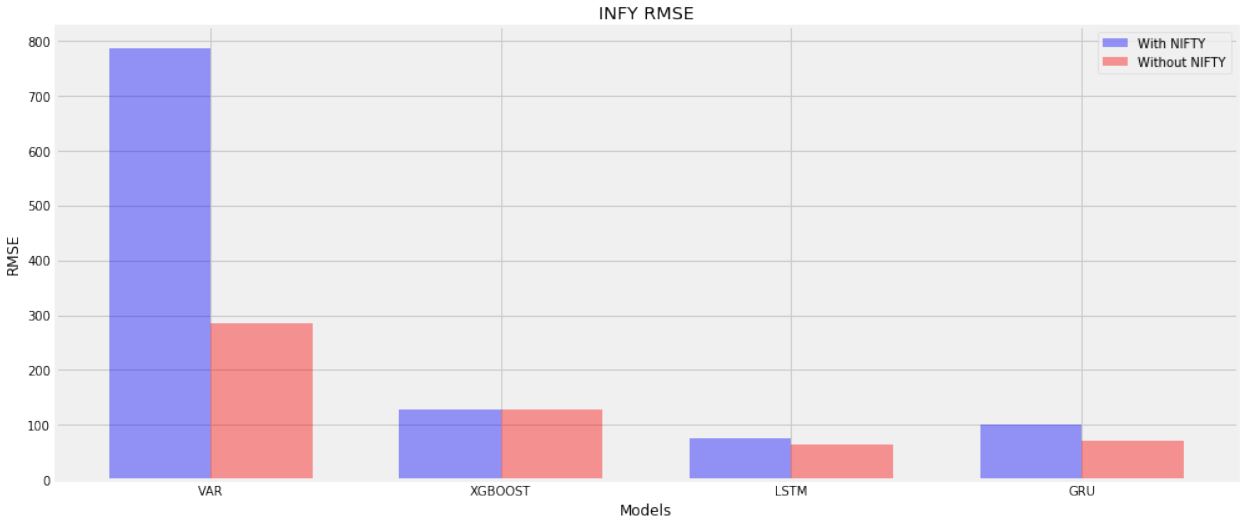
- **Results:**

  As can be observed from the above plots, it is difficult to conclude whether NIFTY has any significant effect on the predictions. Hence it is necessary to chart the RMSE values for each model with and without NIFTY as the exogenous parameter for each company stock. The resulting figures are as follows:

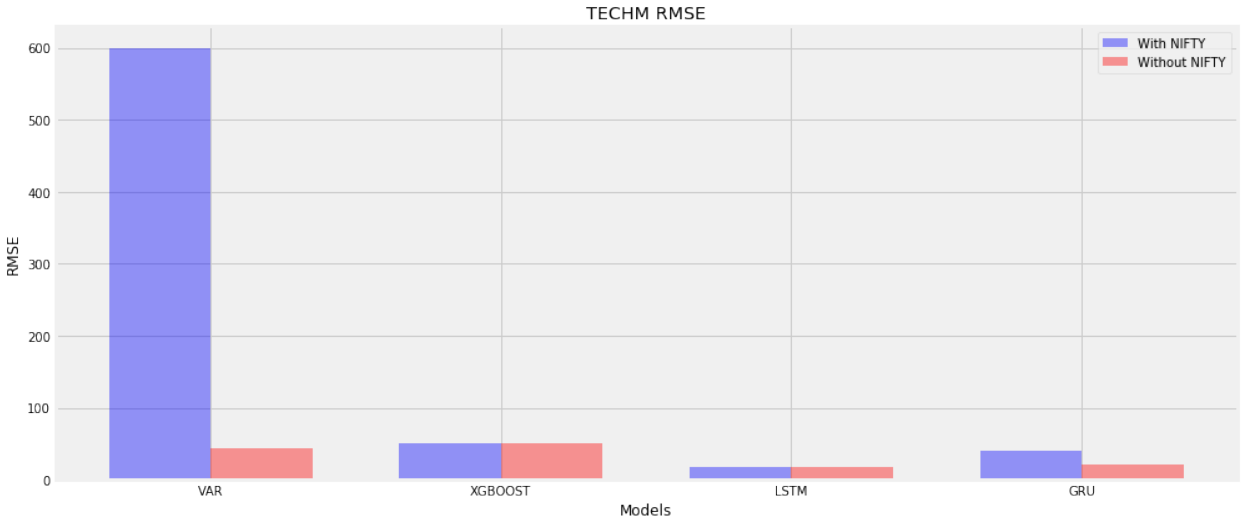
  – TCS RMSE chart:



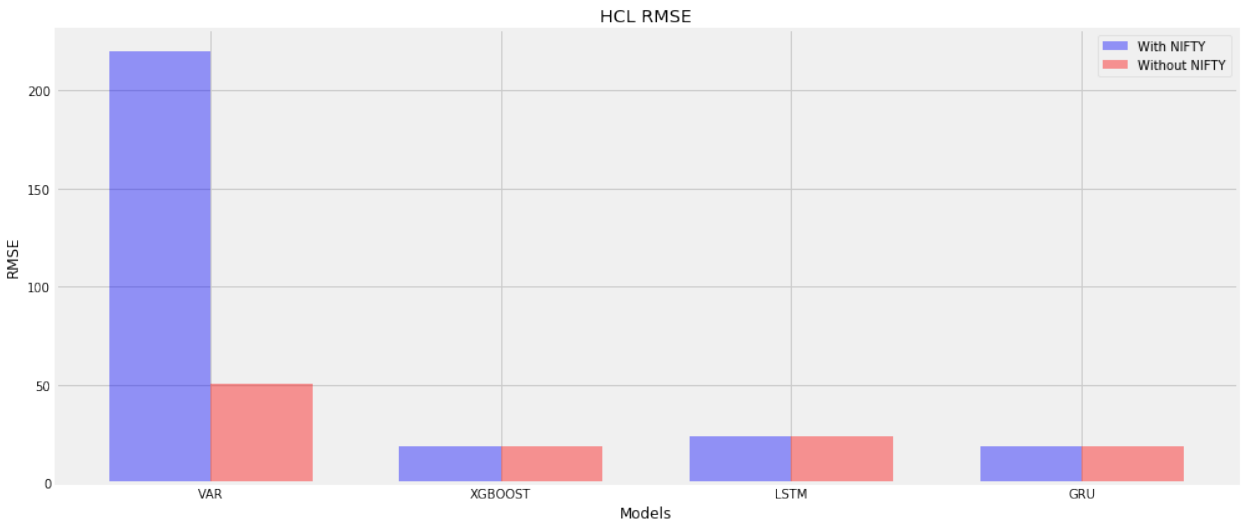
– INFY RMSE chart:



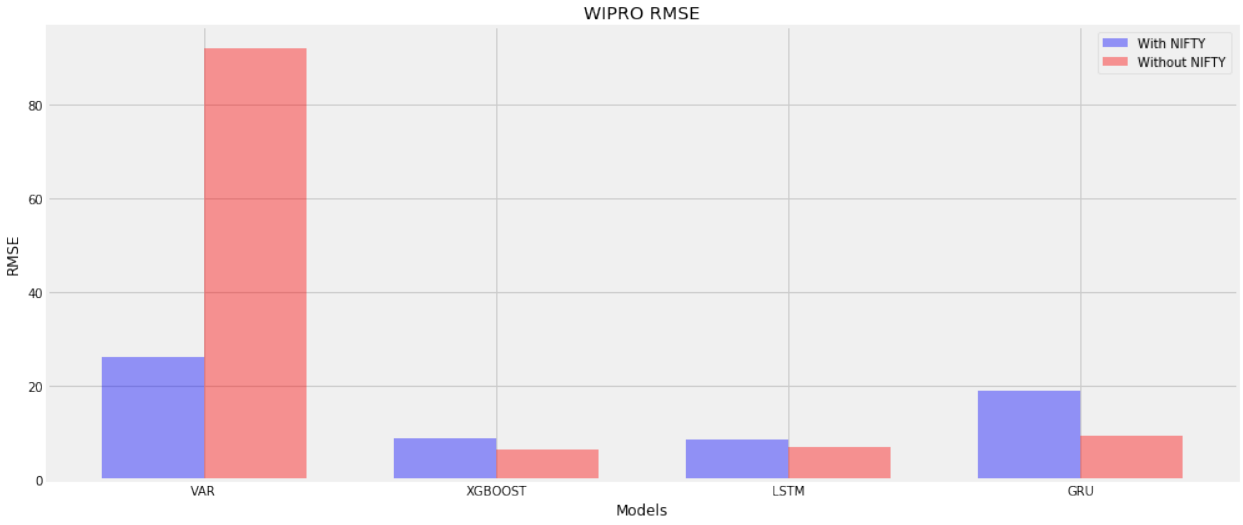
– TECHM RMSE chart:



– HCL RMSE chart:



– WIPRO RMSE chart:



As can be surmised from the above charts, the experimental models almost always perform worse when using NIFTY as an exogenous variable.

5 Conclusions and Future Scope

The following conclusions can be arrived at after looking at the results available:

- In case of the VAR model, it an be observed that NIFTY as an exogenous variable consistently and significantly worsens the models performance except in case of WIPRO where we see an enormous improvement in performance. The reason found upon further investigation, was that the WIPRO stock is an idle one and hence changes happen almost always in tandem with the economy as a whole which is represented by NIFTY

- In case of the other three models, namely XGBOOST, LSTM and GRU, it is observed from our results that they almost always perform worse when using NIFTY as an exogenous variable. The discrepancy observed with VAR, as mentioned in the previous point doesn't happen with the other three because they are much better equipped to capture the relative changes in economy as well as the individual company stocks.

Although we can conclude that our hypothesis of NIFTY being a good exogenous variable for stock prediction of Indian IT stocks was erroneous, we cannot say the same about a more general hypothesis which says that there exists a set of exogenous variables that can significantly improve the accuracy of individual stock prediction. Keeping this in mind, the following can be considered as areas which show scope for future enhancements:

- Identify other exogenous variables other than NIFTY such as commodities values and add them to the feature set.

- Exhaustively try different networks for Deep Learning models while tuning the hyper-parameters number of epochs, dropout probability etc.

- Experiment with stocks that depend more on overall market like BSE stocks and small-cap which will make the experimental setup more diverse than that of the one using just technology stocks.

- Incorporate early-stopping for deep learning models in order to increase accuracy.

- Experiment with a longer time-frame so as to see if there is another order of seasonality that was missed. This might reflect long-term external stimuli like change of Govt., elections, declaration of budget etc.

6 Acknowledgment

The project group members would like to thank the following for the successful completion of the project:

- Professor Janardhan Rao (Jana) Doppa for giving them the opportunity to work on the proposed topic and for his help while choosing the experimental models.

- Stackoverflow and Github for providing the necessary guidance while faced with implementation issues.

The workload of the project was divided equally among the group members which is reflected in the commit history of the Project Gitlab Code Repository available at Project Commit History.

References

[1] Durbin, James, and Siem Jan Koopman. 2012.Time Series Analysis by State Space Methods: Second Edition.Oxford University Press.

[2] Hyndman, Rob J., and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2014.

[3] Lutkepohl (2005) New Introduction to Multiple Time Series Analysis.

[4] https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf

[5] http://www.bioinf.jku.at/publications/older/2604.pdf

[6] https://arxiv.org/pdf/1406.1078.pdf