Methodology

1. Price Prediction

For the forecasting of the stock prices we compared the performance of three models i.e Long Short Term Memory(LSTM) model, Facebook Prophet(FbProphet), and the Autoregressive integrated moving average(ARIMA) model. Proof of concept was done on different datasets. For the purpose of this research we have taken historical stock prices data from 2002 to 2020 of companies such as Maruti, Tcs, Nestle, Tata Steel. These companies have shown different trends in the past 18 years which are shown below in the figures.



Figure 1: Trend of Maruti from 2002 to 2020

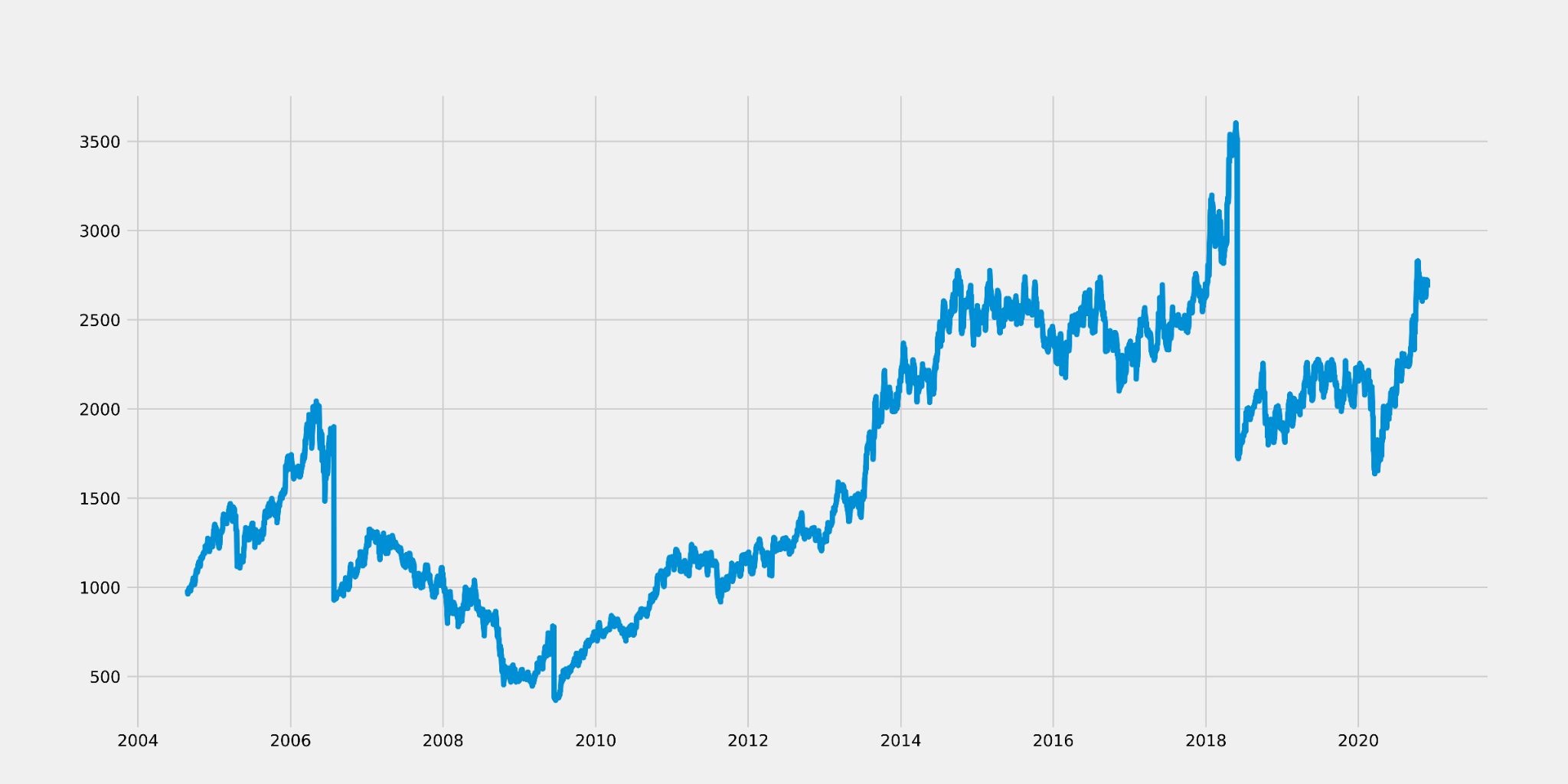


Figure 2: Trend of TCS from 2002 to 2020

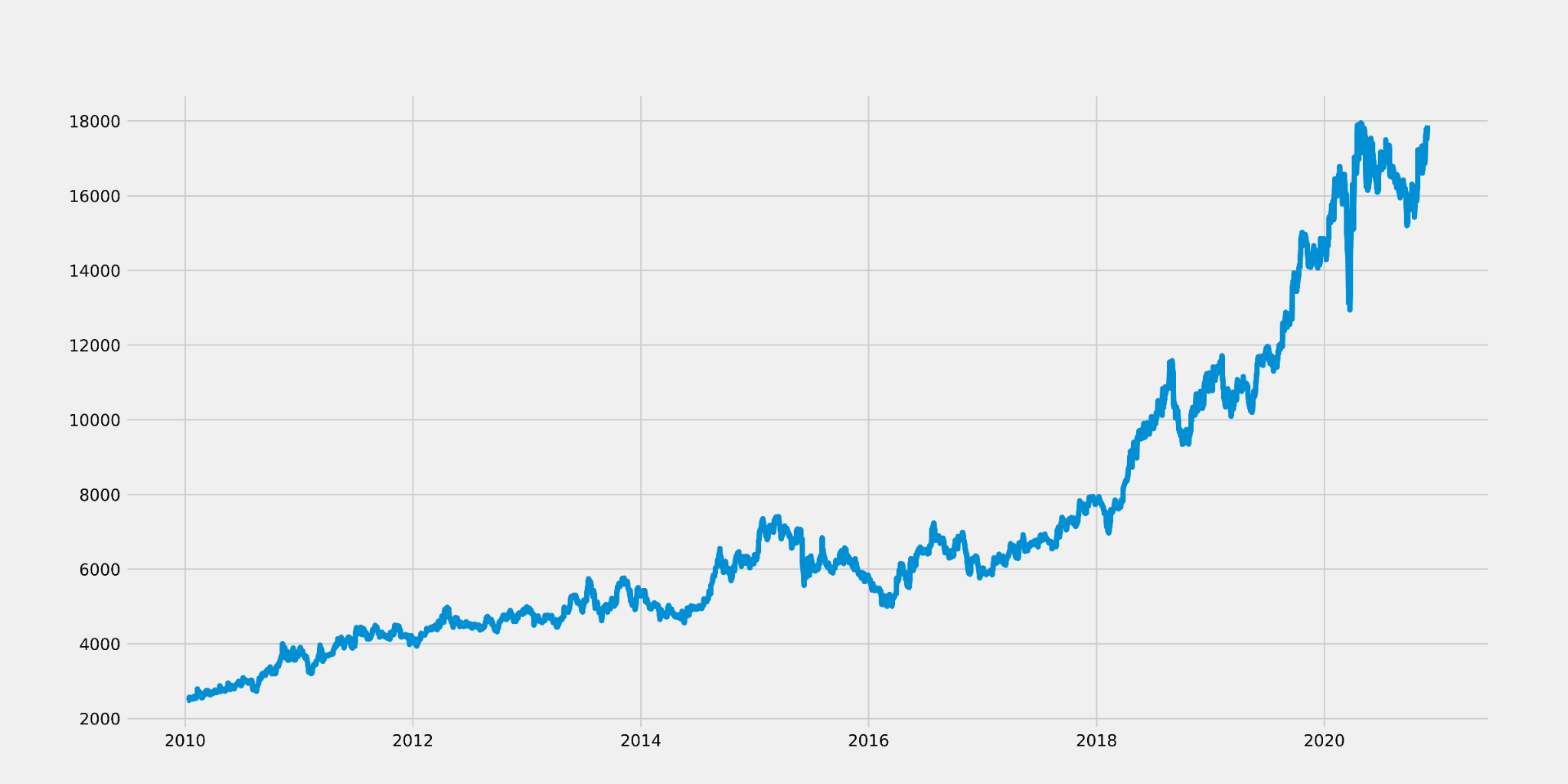


Figure 3: Trend of Nestle from 2002 to 2020

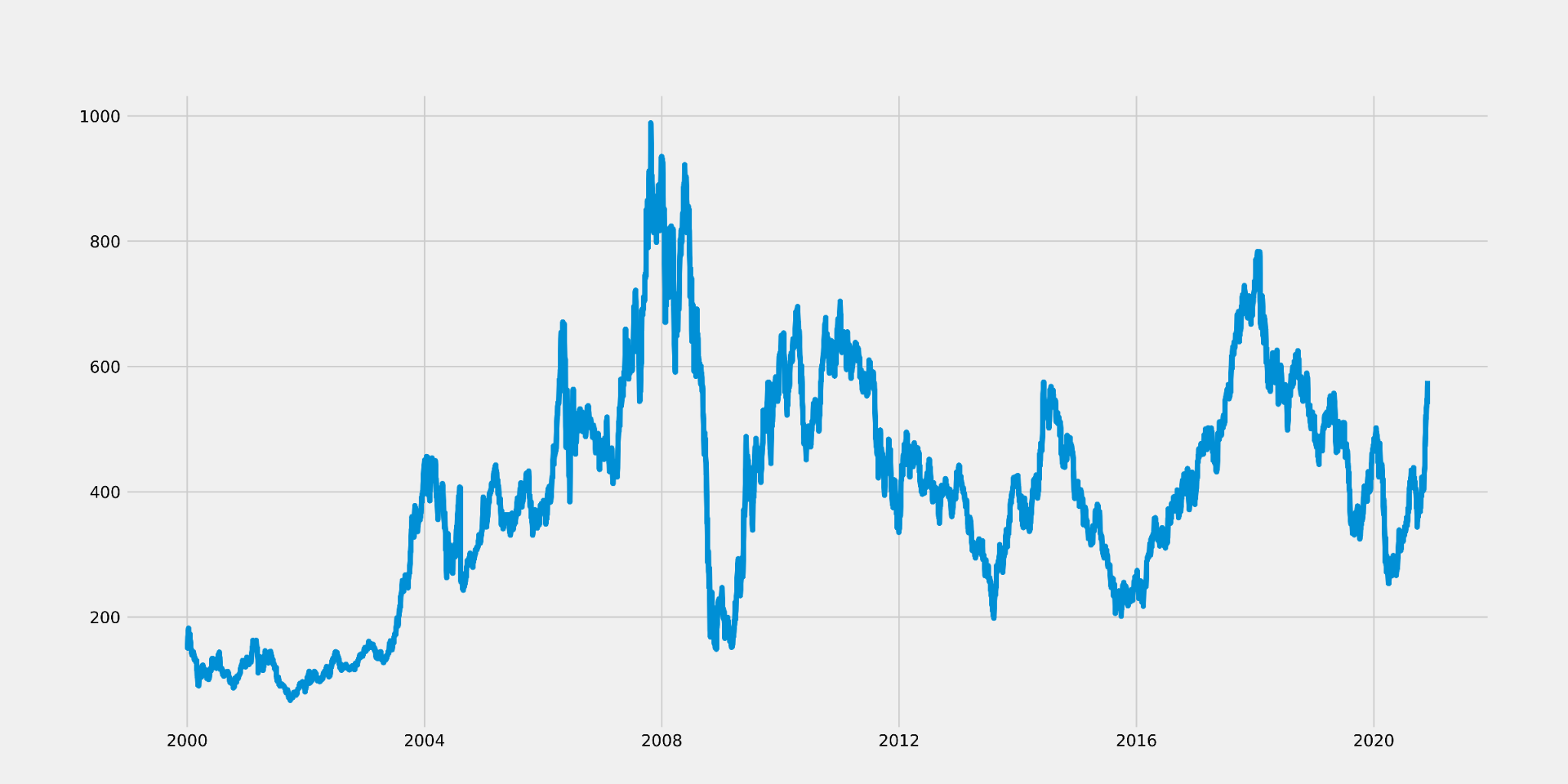


Figure 4: Trend of Tata Steel from 2002 to 2020

* Auto Regressive integrated moving average(ARIMA)

Arima is a statistical model which works great for time series data. It works on two main assumptions i.e because the data series is steady, the mean and variance should not change over time therefore the log transformation or differencing of a series can be used to make it stationary and because arima predicts future values based on previous values, the data supplied as input must be a univariate series.

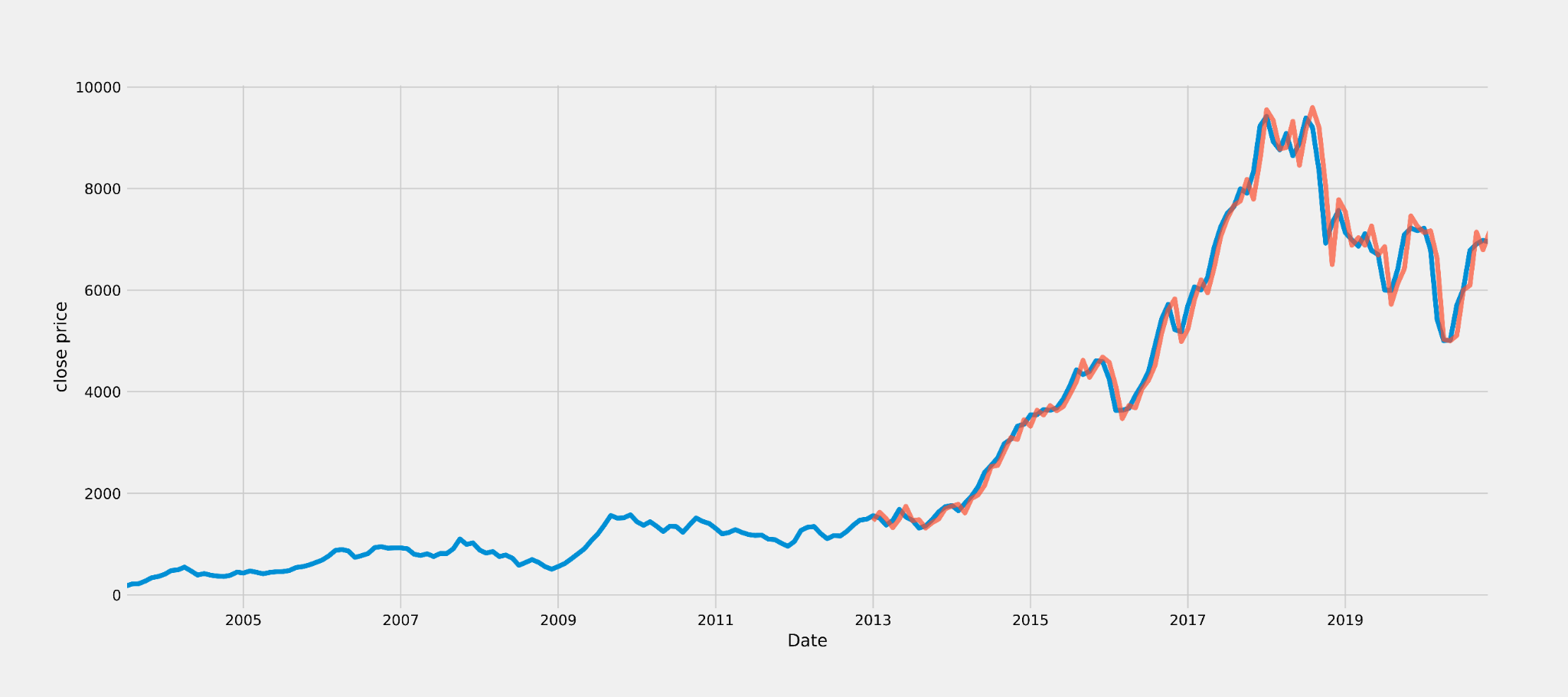


Figure 5: Actual vs Predicted for Maruti using ARIMA

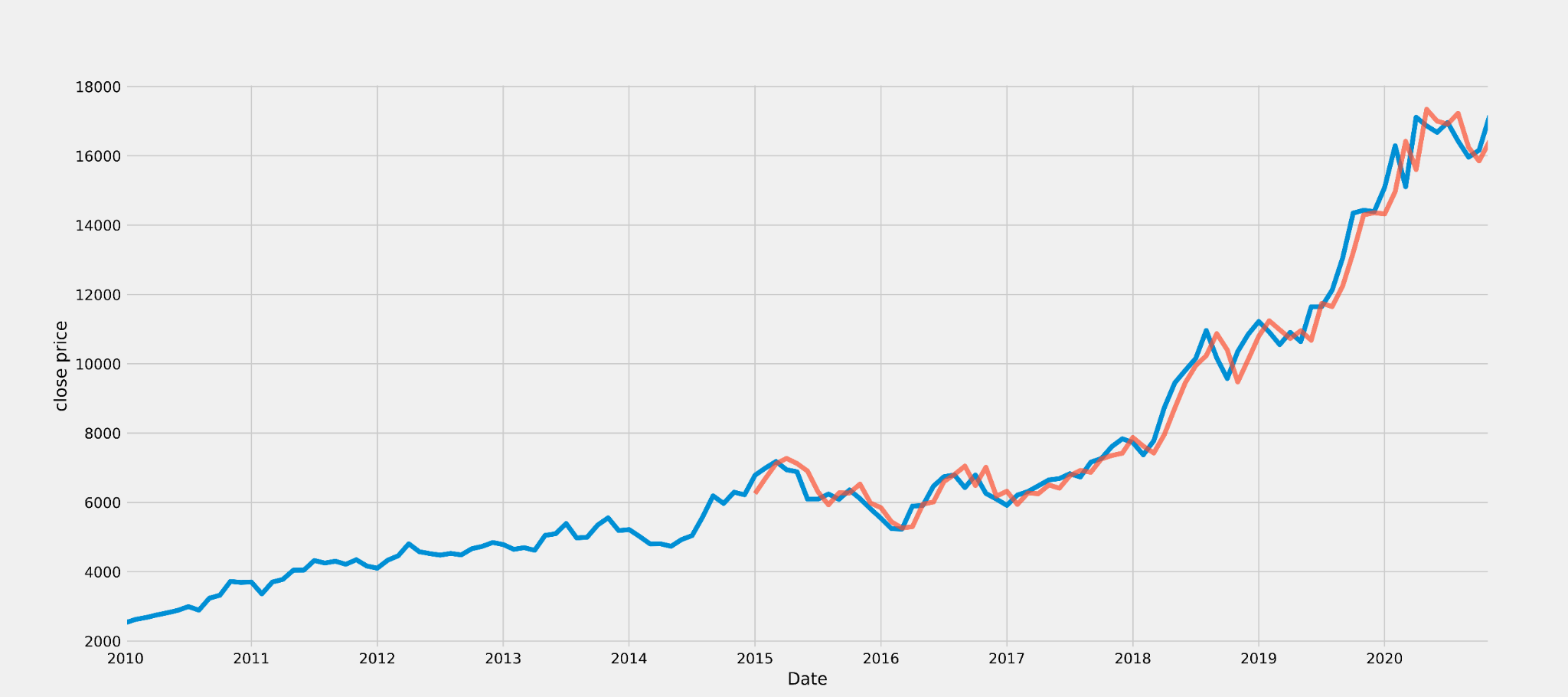


Figure 6: Actual vs Predicted for Nestle using Arima

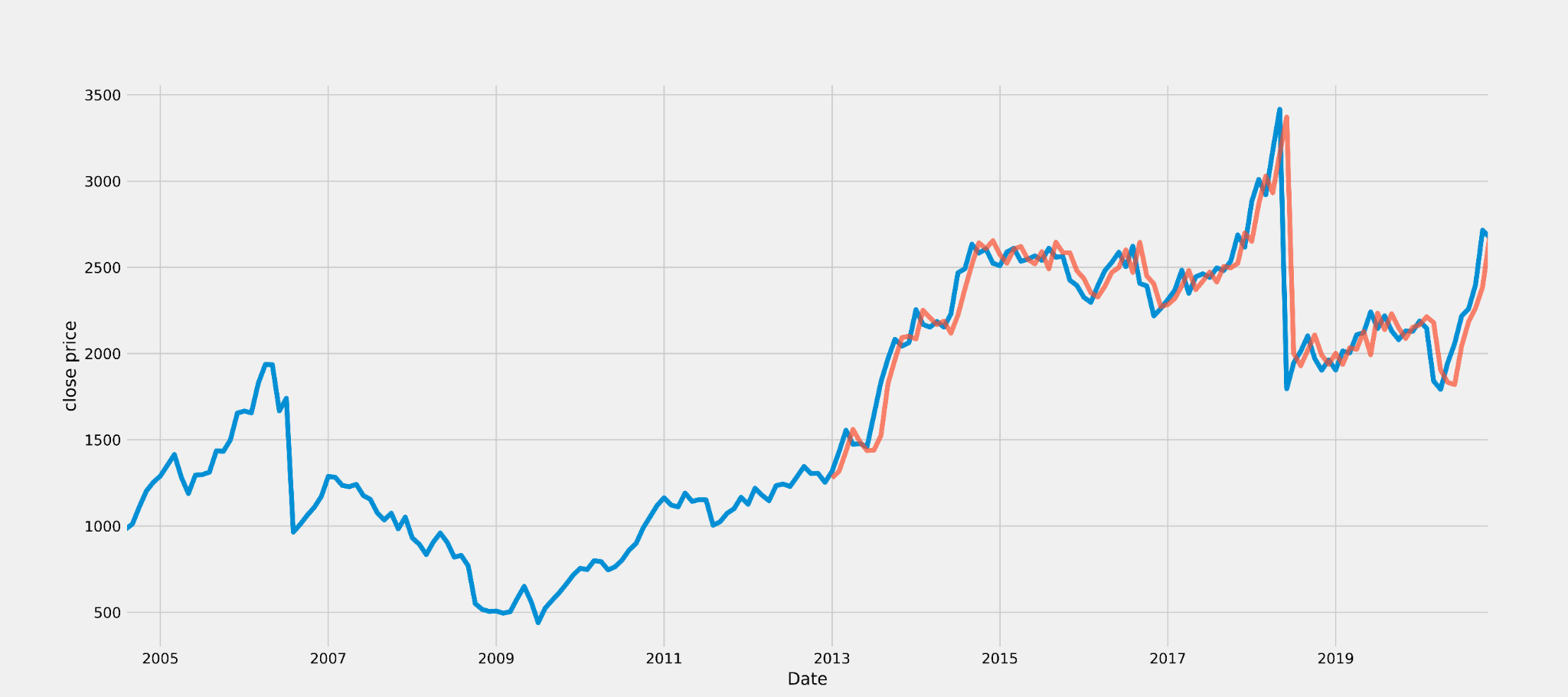


Figure 7: Actual vs Predicted for TCS using Arima

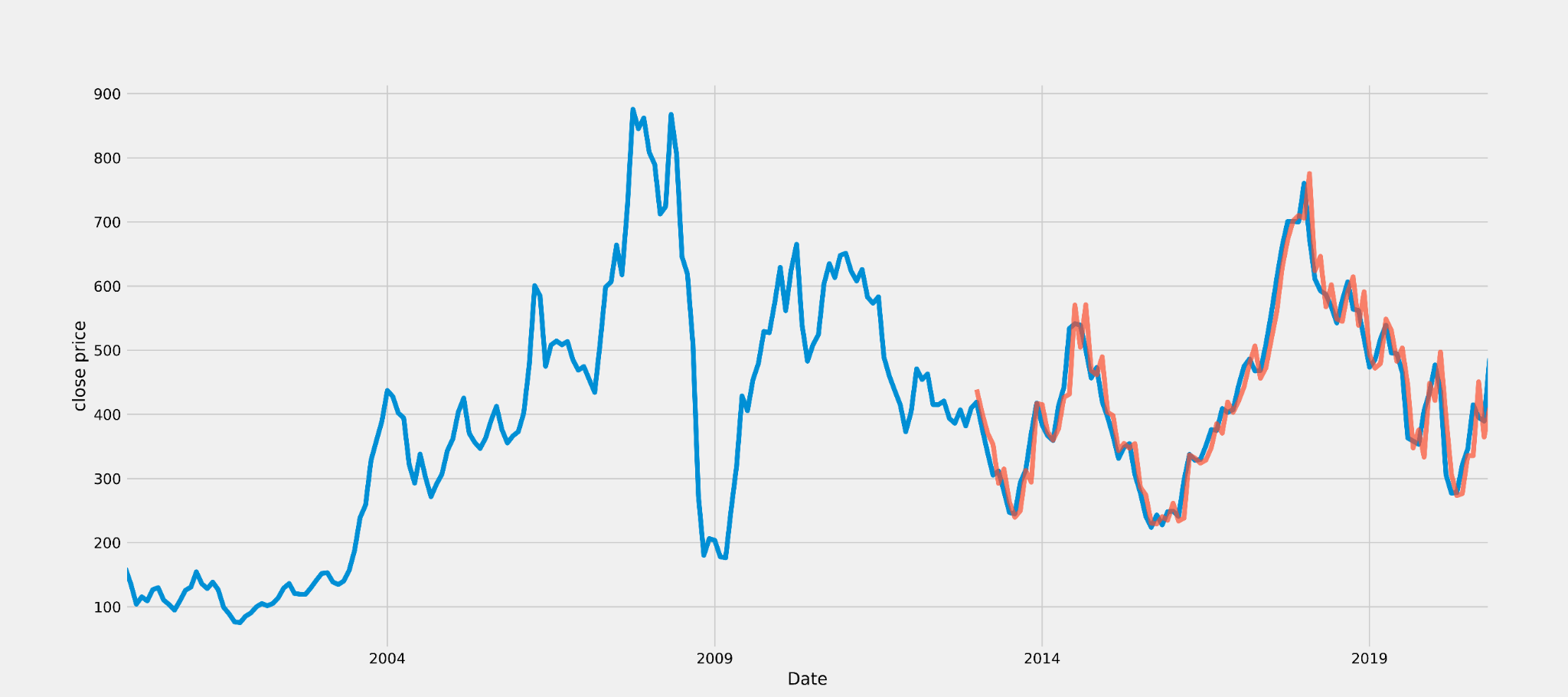


Figure 8: Actual vs Predicted for Tata Steel using Arima

As we can see in the results given in table1, different datasets give us different accuracies. Here we calculate the root mean square error(RMSE) value.

| Stock Name | RMSE value |
| --- | --- |
| Maruti | 12.8 |
| Nestle | 17.7 |
| TCS | 15.4 |
| Tata Steel | 11.1 |

Table 1: Results of ARIMA model

* Facebook Prophet(FbProphet)

There are a variety of approaches that may be used to the stock prediction dataset, but the majority of them need extensive data preparation prior to fitting the model. Prophet is a time series forecasting toolkit created and pioneered by Facebook that requires no data preparation and is incredibly easy to use. Prophet takes a two-column data frame as input.

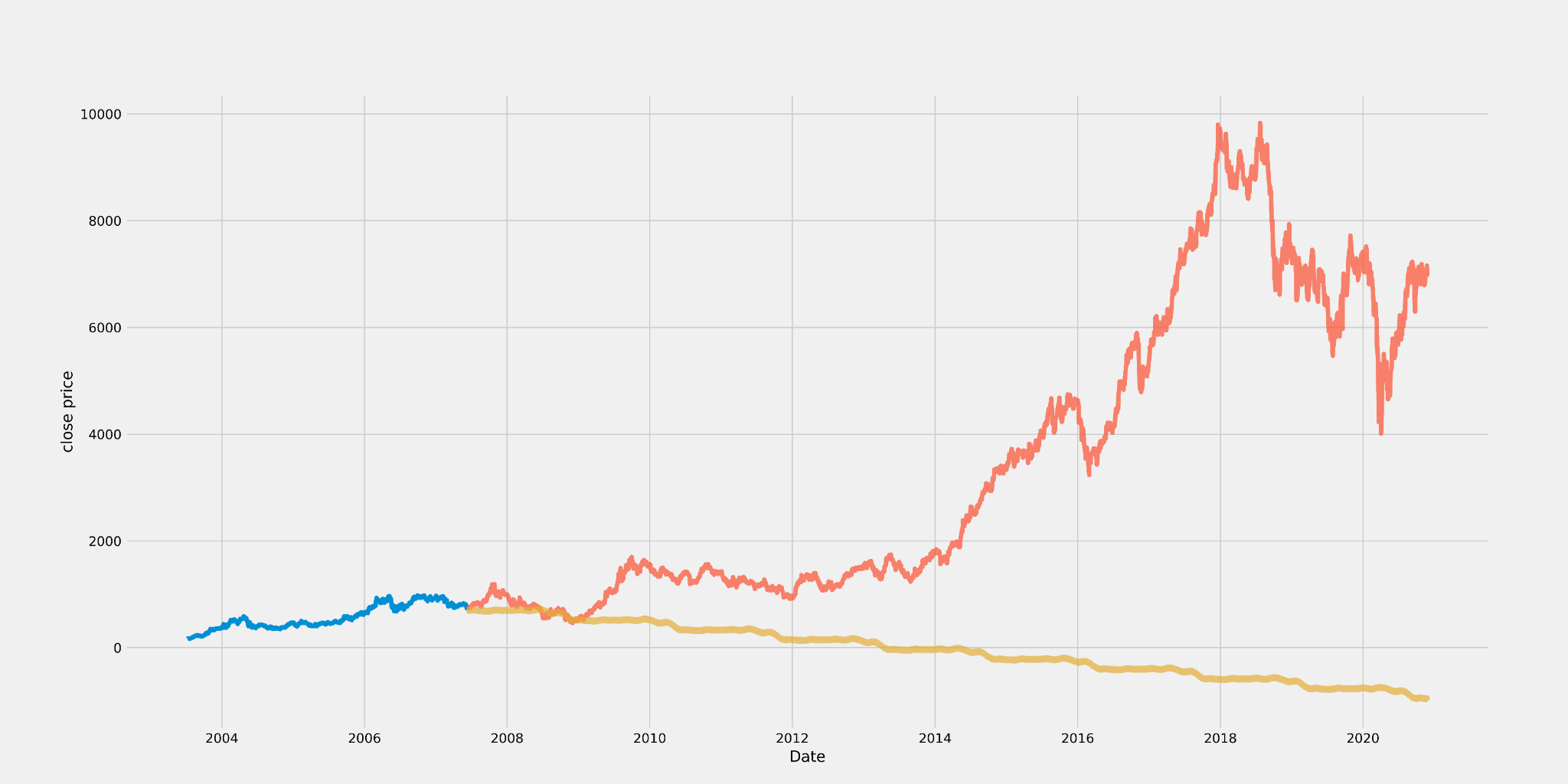


Figure 9: Actual vs Predicted for Maruti using FbProphet

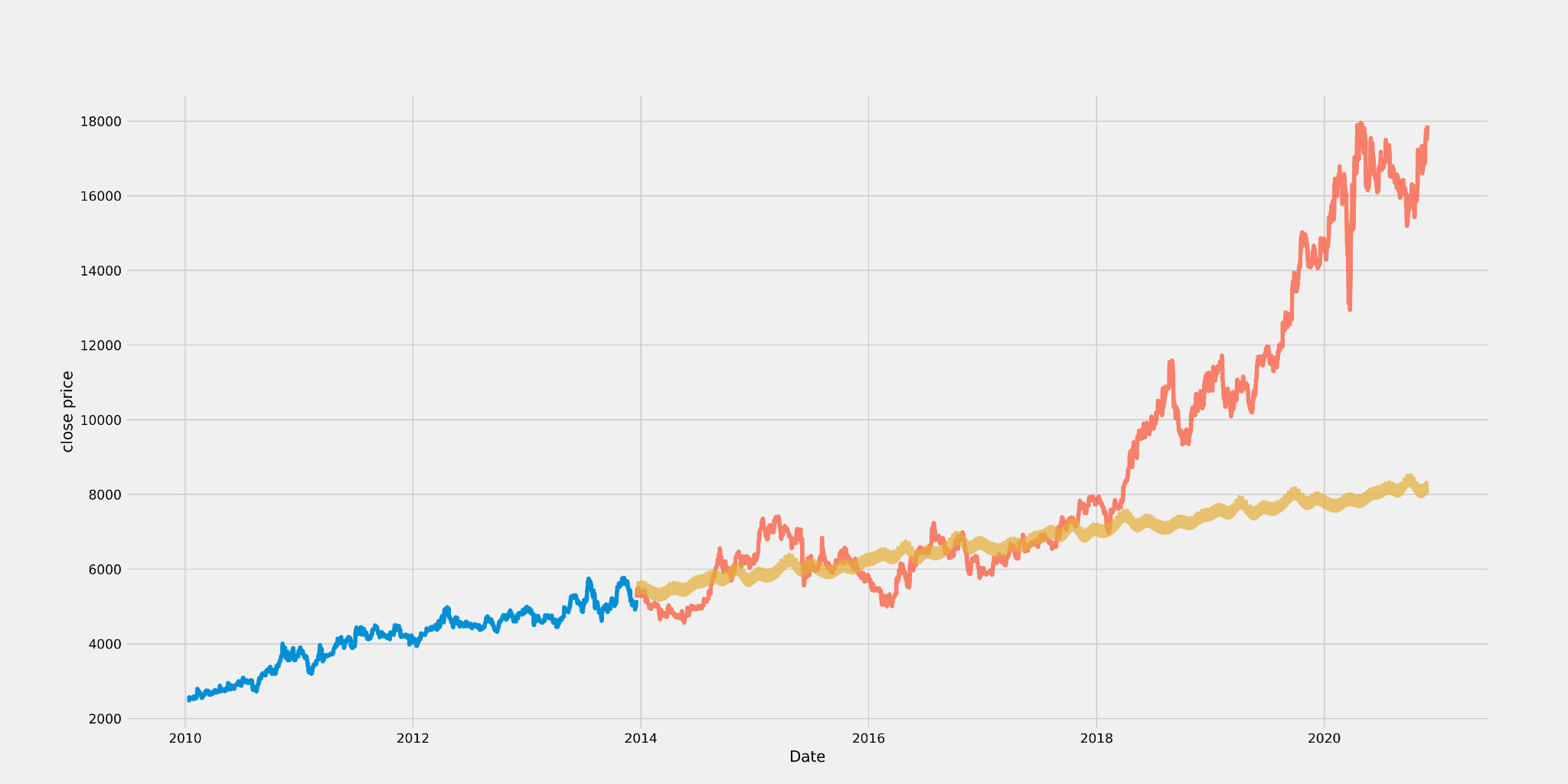


Figure 10: Actual vs Predicted for Nestle using FbProphet

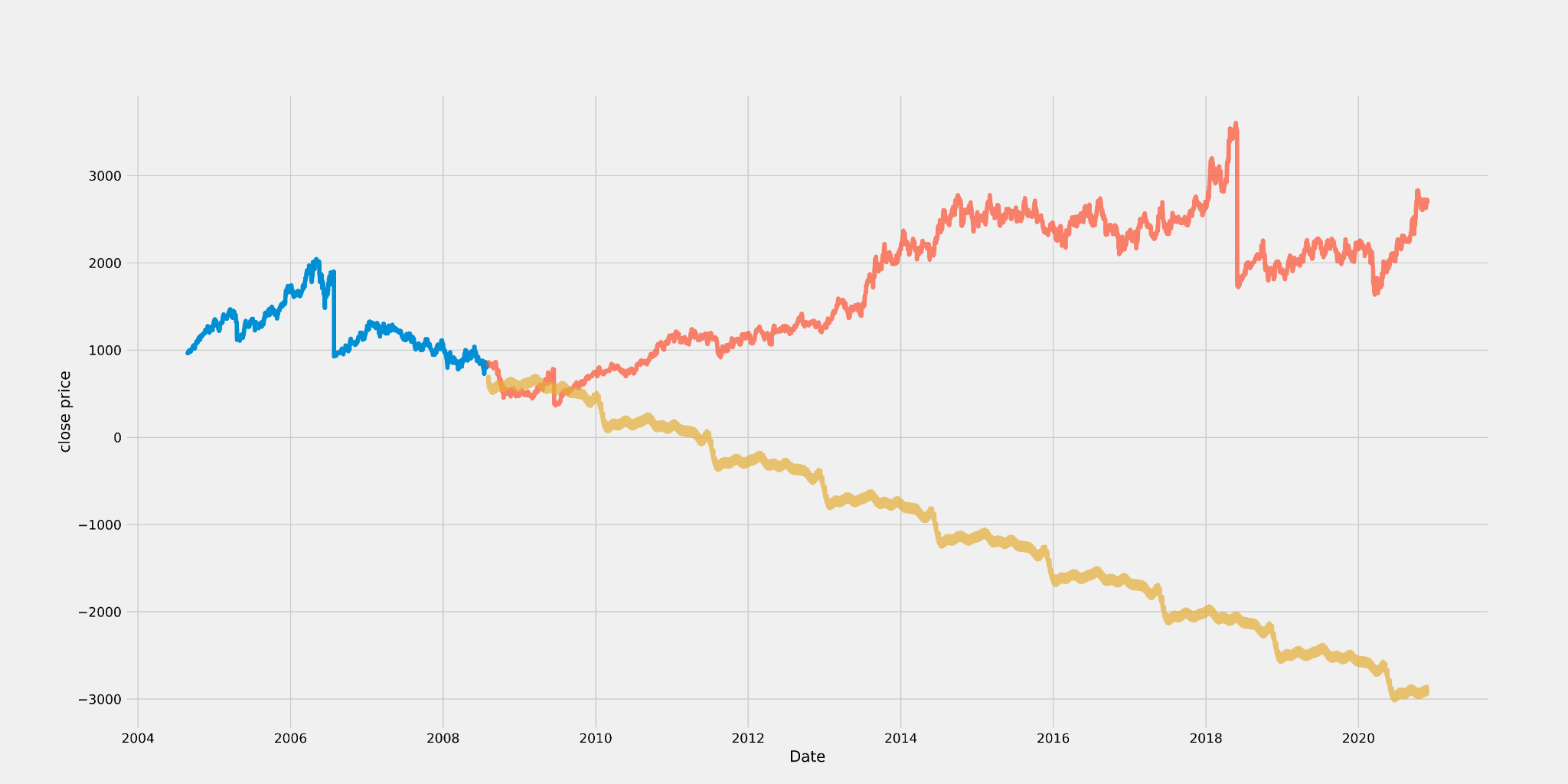


Figure 11: Actual vs Predicted for TCS using FbProphet

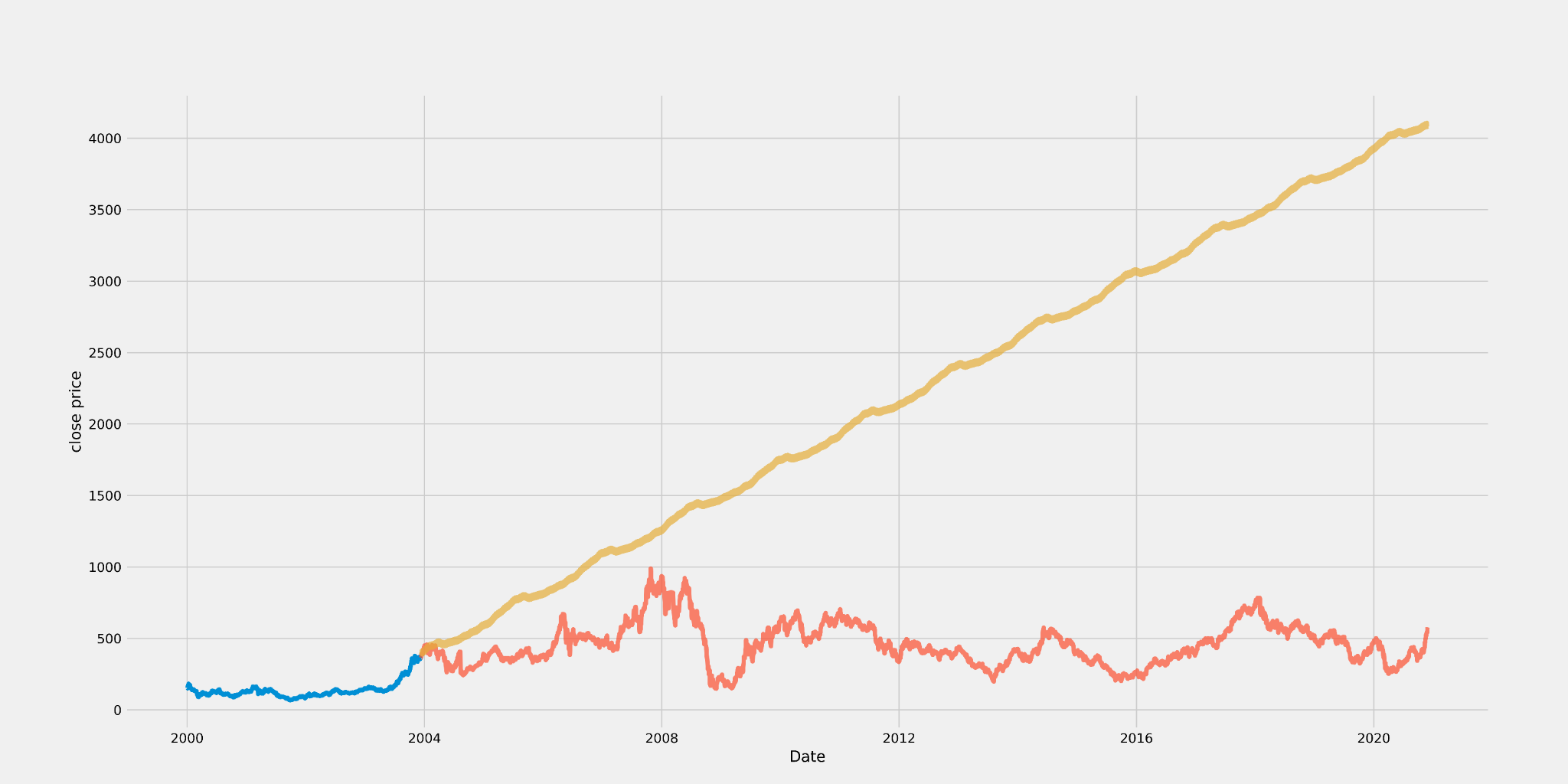


Figure 12: Actual vs Predicted for Tata Steel using FbProphet

As we can see in the results given in table 2, different datasets give us different accuracies. Here we calculate the root mean square error(RMSE) value.

| Stock Name | RMSE value |
| --- | --- |
| Maruti | 98.2 |
| Nestle | 70.2 |
| TCS | 97.7 |
| Tata Steel | 101.5 |

Table 2: Results of FbProphet

* Long short term memory(LSTM)

For sequence prediction issues, LSTMs are frequently used and have shown to be highly successful. They function so effectively because LSTM can remember vital information from the past while forgetting irrelevant information. There are three gates in an LSTM: an input gate, a forget gate, and an output gate.

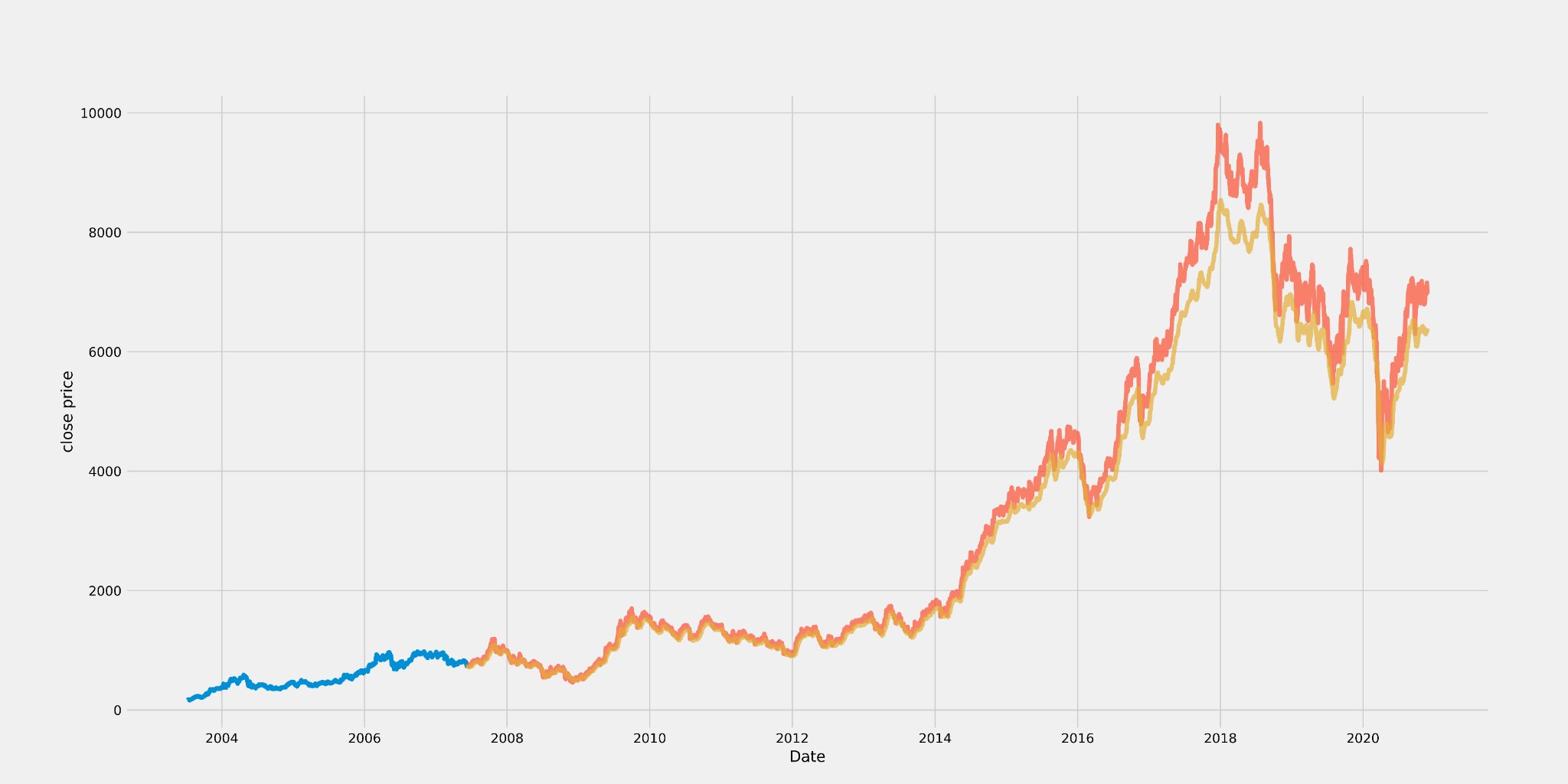


Figure 13: Actual vs Predicted for Maruti using LSTM

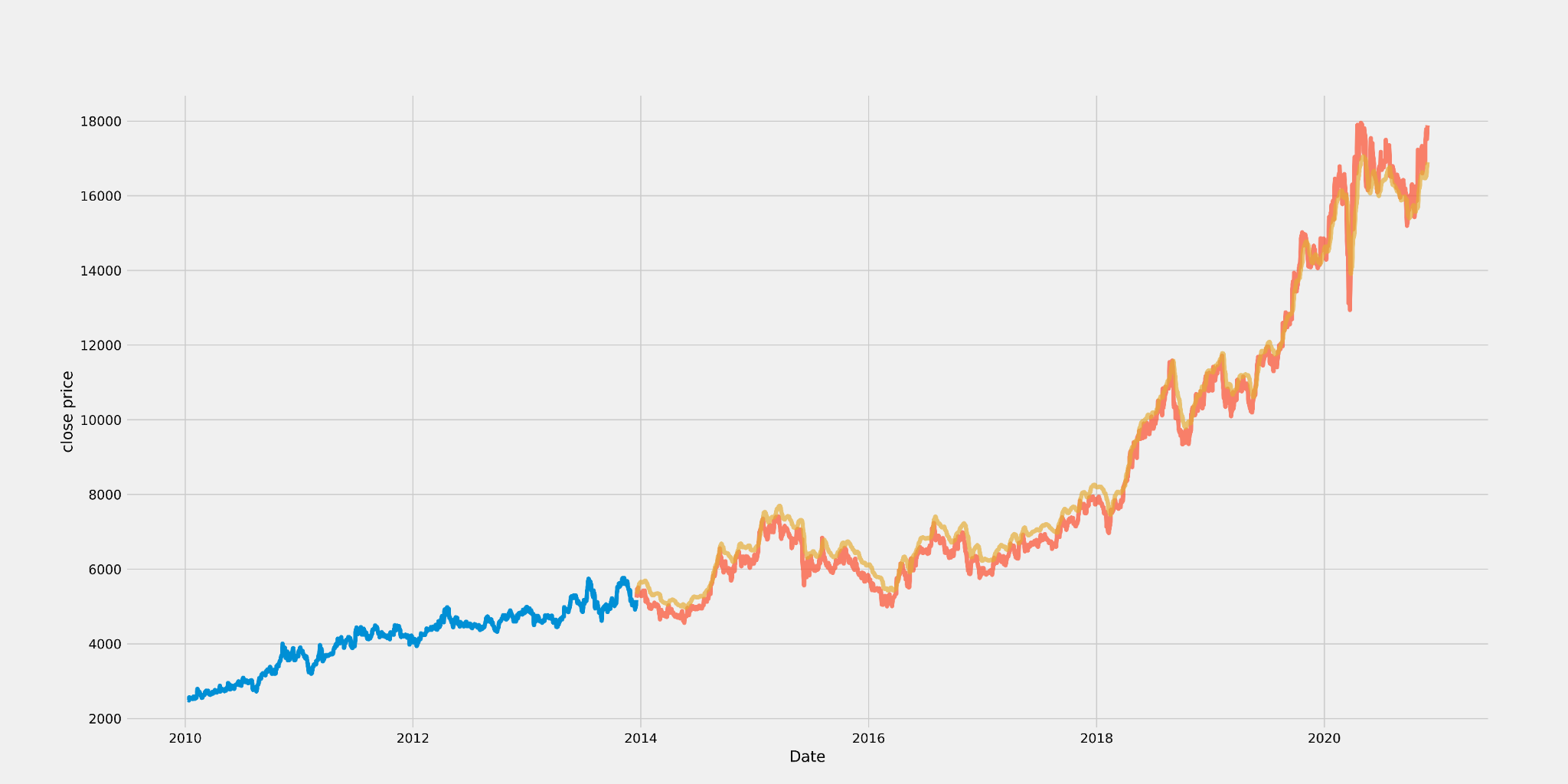


Figure 14: Actual vs Predicted for Nestle using LSTM

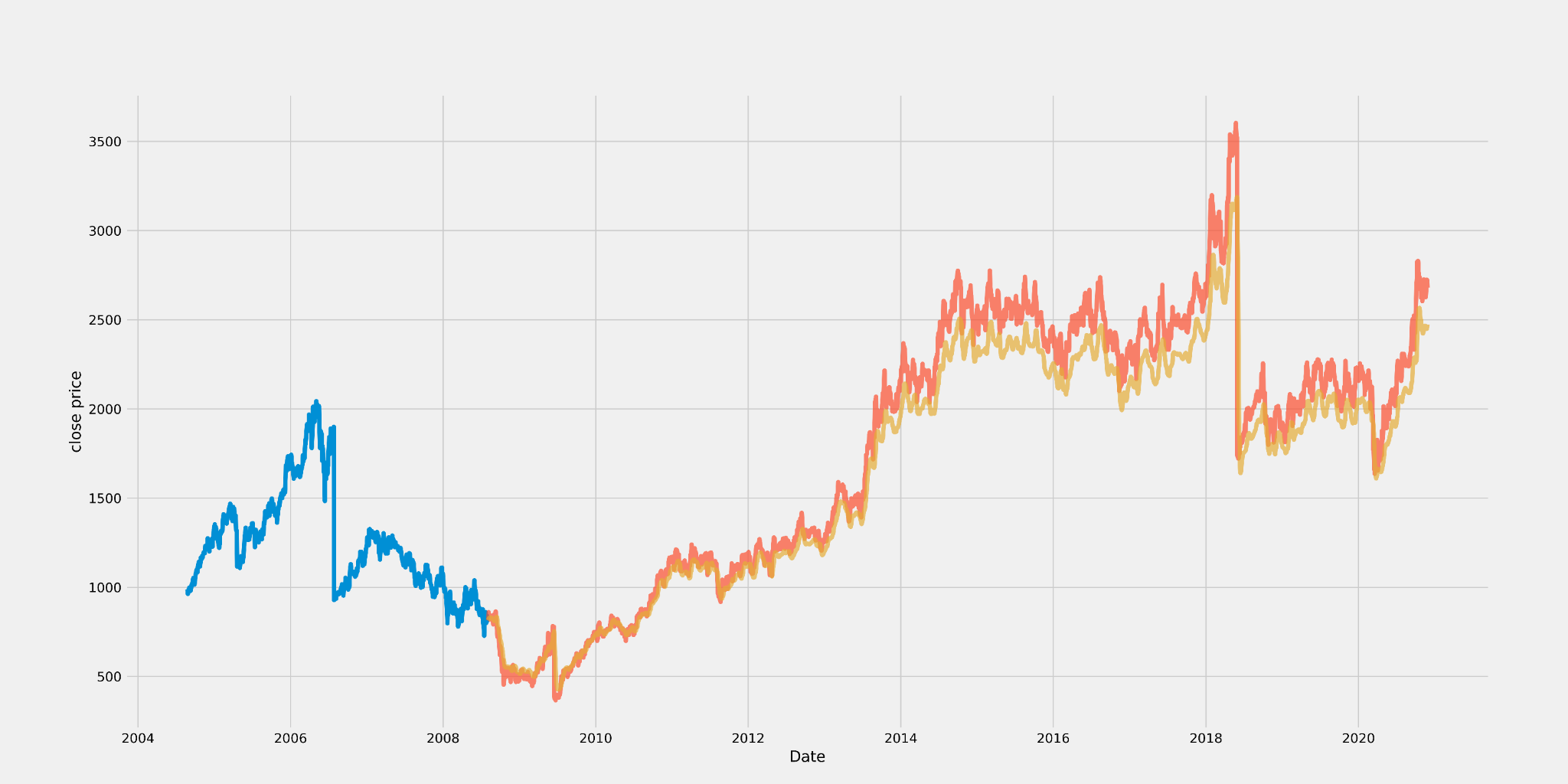


Figure 15: Actual vs Predicted for TCS using LSTM

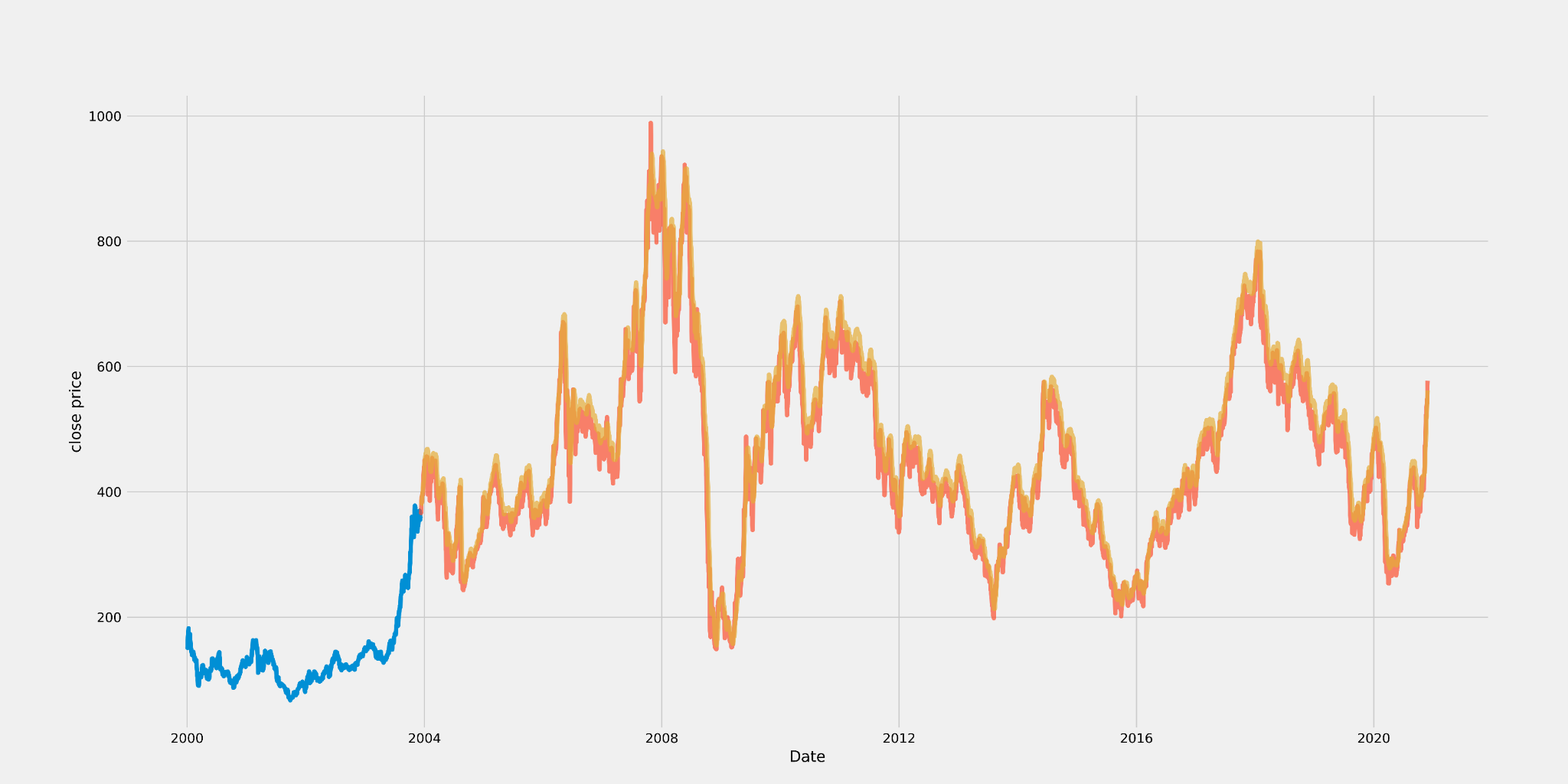


Figure 16: Actual vs Predicted for Tata Steel using LSTM

As we can see in the results given in table3, different datasets give us different accuracies. Here we calculate the root mean square error(RMSE) value.

| Stock Name | RMSE value |
| --- | --- |
| Maruti | 14.8 |
| Nestle | 10.2 |
| TCS | 11.8 |
| Tata Steel | 9.7 |

Table 3: Results of LSTM

1. Sentiment Analysis

For the purpose of training and testing the sentiment analysis model we used a dataset which contains 5790 stock related tweets with 3500 positive and 2290 negative tweets. After all the data preprocessing steps we trained and tested 6 different models - LSTM, Naive Bayes, Random Forest, KNN, Logistic Regression, and MLP classifier. The table 4 shows the different accuracies that were calculated. The accuracies were calculated using the confusion matrix.

| Models | Accuracy(%) |
| --- | --- |
| LSTM | 78.084 |
| Naive Bayes | 83.22 |
| Random Forest | 78.86 |
| Logistic Regression | 79.03 |
| KNN | 67.04 |
| MLP | 77.39 |

Table 4: Results of different Sentiment Analysis models