Results

All the experiments were broadly conducted under two different categories, the first category was without acknowledging the data as time series data and the other was considering it as time series data.

For the first category, features were generated using statistical formulas of technical indicators like Relative Strength Index, Moving Average etc. After generating 100 such features, 25 of them were selected using Recursive Feature Elimination(RFE) technique. Using those 25 as features the trend was predicted and the results were:-

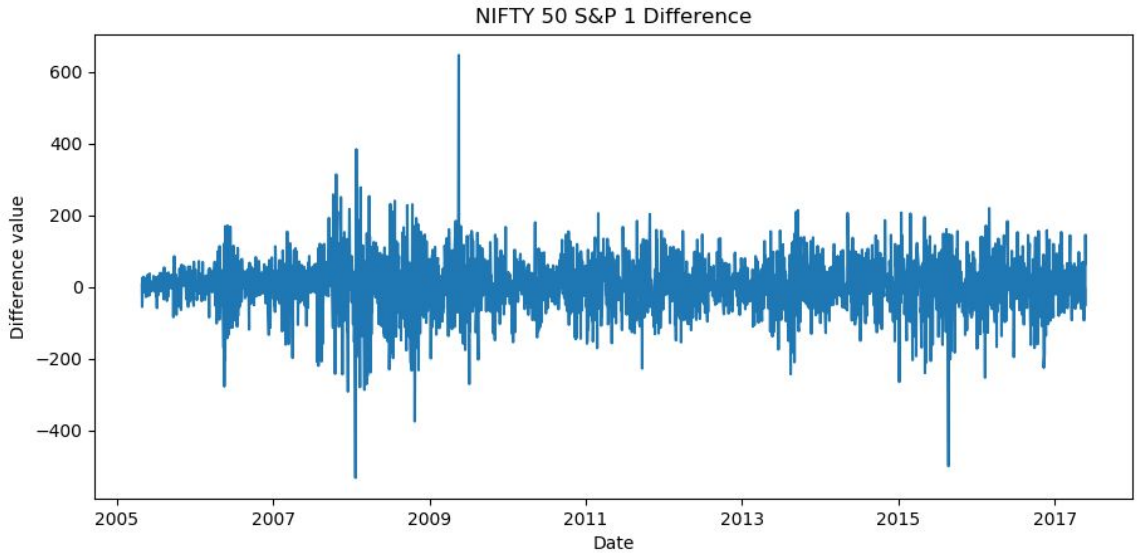
Model	**Accuracy**
Lasso	55.3%
Ridge	54.7%
Naive Bayes	52.0%
K Nearest Neighbour	53.9%
Linear Discriminant Analysis	56.3%
Support Vector Machine	61.7%
Random Forest	63.4%

Later, I tried ensemble models like using Random Forest and Naive Bayes in a single voting classifier. But, the results did not improve. However, the confidence of true positives was high and there were lesser false positives.

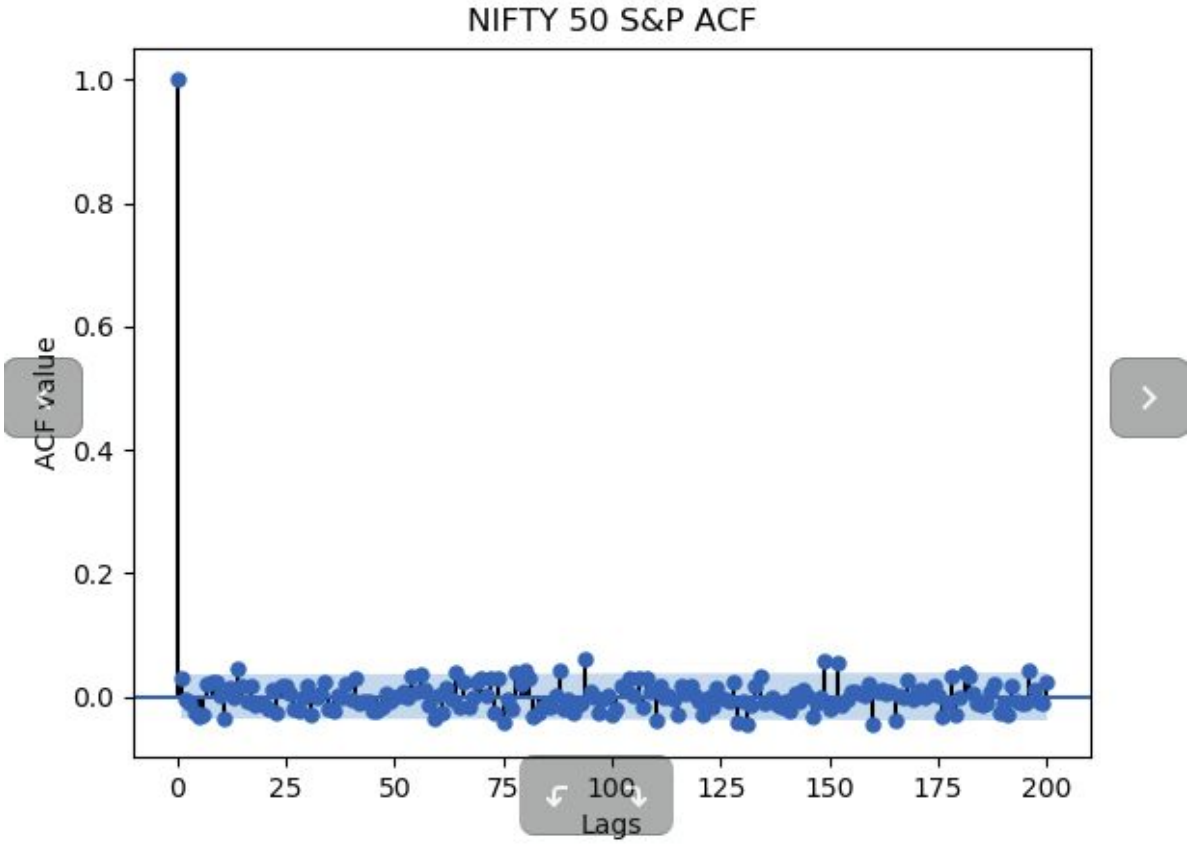
In the same category, I tried several neural network models. Finally, one model with two layers of LSTM and one output layer Dense layer produced good results. The result went even better than Random Forest with rolling window and hyperparameter tuning for some of the data like JetAirways the accuracy went up to 67% for some time interval.

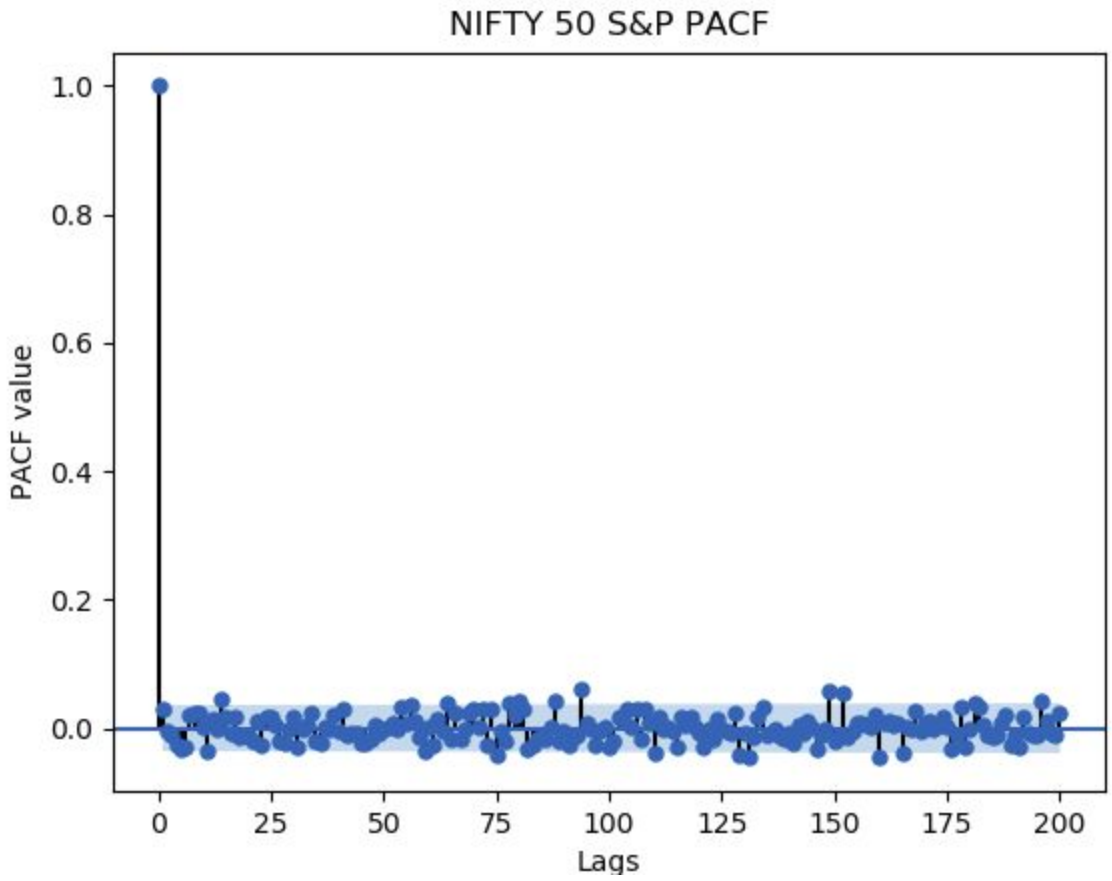
For the second category, the data was acknowledged to be time series data. Instead of using the features generated, the closing/opening price was used directly as a feature. But, the data is non-stationary as the mean is not zero and variance is also not constant. So, to make data stationery, we need to take the first or second difference. After the first difference itself, the data seems to be stationary with zero mean and constant variance with reasonable

fluctuation(considering the festivals). Dicky-fuller test also suggests that the data plotted below is stationary.



To apply ARIMA we need three parameters (p, d, q). P is for auto-regressive(AR) part, Q for Moving Average(MA) part and D for Integrated(I). So, as we can see the data is stationary after first difference we could use d=1. Similarly, the below ACF and PACF plot suggests P=1 and Q=1.(*I would give a detailed explanation in final report.*)





So, with parameters (p=1, d=1, q=1), the results are good for Nifty50 over large span of time and it is more than 66%. For some smaller time intervals the result went upto 75%.

The good result in ARIMA is clearly because of accounting the fact that our data is time series data. Making it stationary was a good idea because with zero mean and constant variance, prediction become much easier to replicate over times as the behaviour of data repeated to some extent.