TIME SERIES ANALYSIS ON ICICI BANK STOCKS REPORT

Problem statement

The goal is to forecast the Volume Weighted Average Price (VWAP) of Nifty stocks using historical trading data. Accurate prediction of VWAP is critical for traders and investors to make informed decisions, optimize trade execution, and assess market trends. The challenge lies in capturing complex temporal patterns, seasonality, and the influence of multiple exogenous features such as price, volume, turnover, and trade activity. The project explores both traditional time series models (ARIMA, SARIMA, AutoRegressor) and deep learning models (LSTM, RNN) to identify the most effective approach for reliable short-term and long-term VWAP forecasting.

Dataset Description-

https://www.kaggle.com/rohanrao/nifty50-stock-market-data

1. Series: Here EQ stands for equity series of stock market.
2. Prev Close: The closing price of the stock for the day before.
3. Open,High, Low, Last, Close: The opening price, highest price, lowest price, last price and closing price of ICICI shares on the current day.
4. VWAP: Volume Weighted Average Price,the target variable to predict. VWAP is a trading benchmark used by traders that gives the average price the stock has traded at throughout the day, based on both volume and price.
5. Volume: Volume of shares traded on the current day.
6. Turnover: It is a measure of stock liquidity calculated by dividing the total number of shares traded over a period by the average number of shares outstanding for the period.
7. Trades: total number of trades on the current day.
8. Deliverable Volume: is the quantity of shares which actually move from one set of people to another set of people.
9. Deliverable(%): Deliverable volume in percentage.

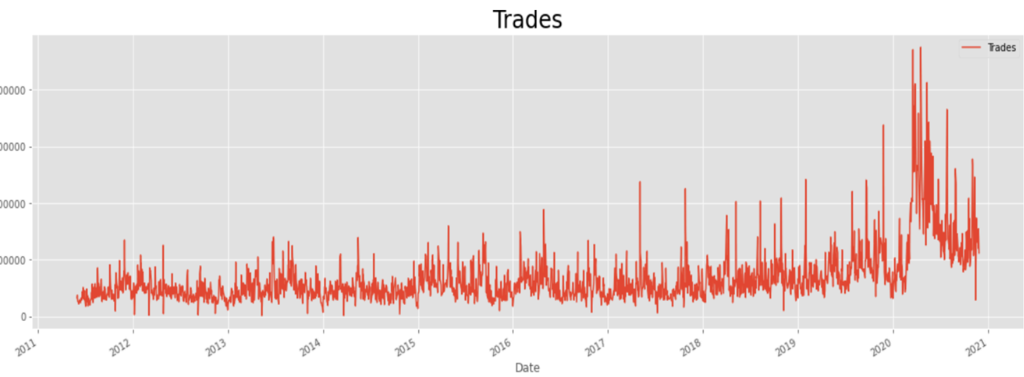
Libraries Used-

1. Statsmodels
2. Keras
3. Numpy
4. Pandas
5. Sci-kit learn
6. Sci-py
7. pmdarima
8. Matplotlib
9. Seaborn

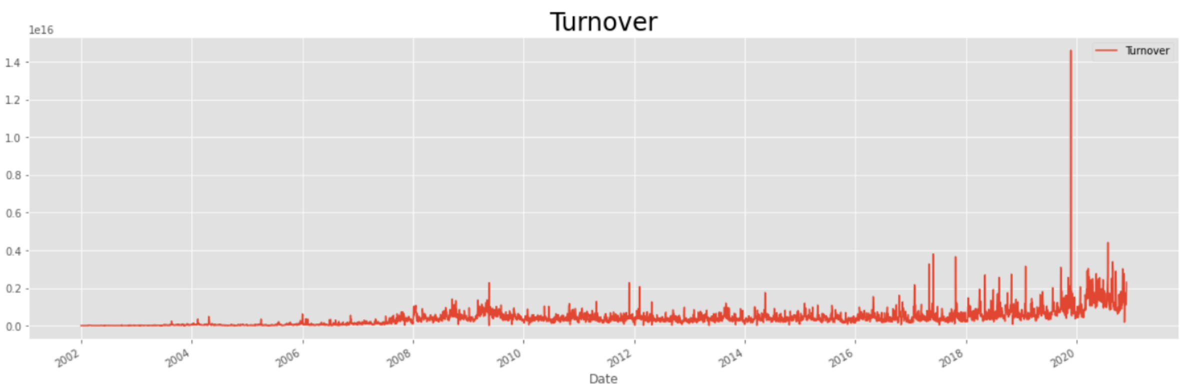
Exploratory Data Analysis (EDA)

1. (5204, 14) is the shape of the data
2. Percentage of missing trade values = 54.77
3. Percentage of missing Deliverable Volume values = 9.93
4. Percentage of missing % Deliverable values = 9.93
5. More than 50% trade data is missing, while only 10% each of deliverable volume and deliverable% is missing. We can drop the rows where deliverable volume is missing. For trade data, we will visualize it to understand the best statistic for imputation.

Visualizing Trends in Trades Column :-

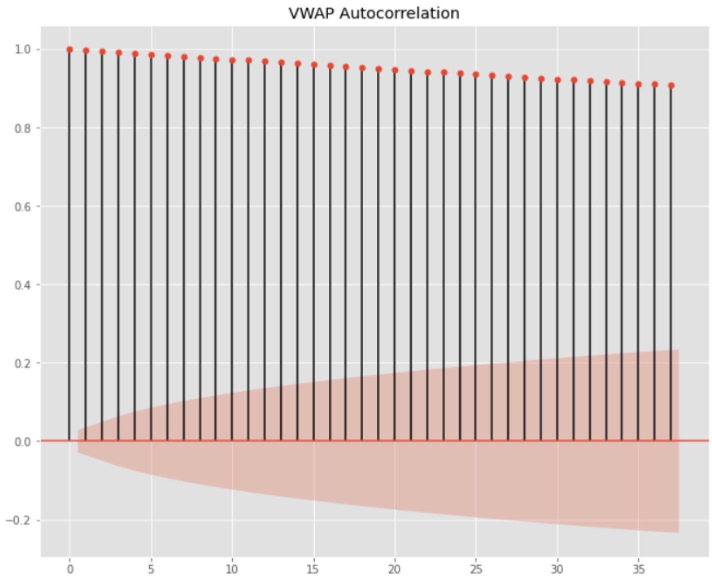


Visualizing Trends in Trades Column :-

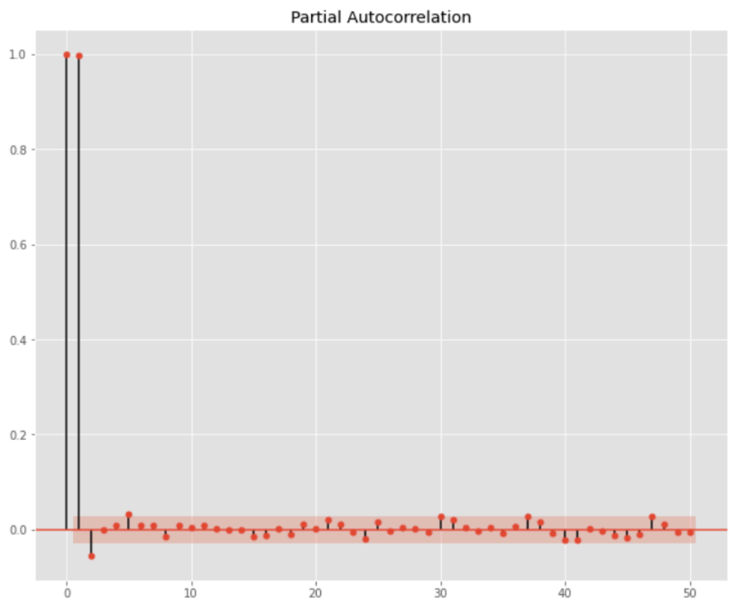


Autocorrelation plot

Autocorrelation - The autocorrelation function (ACF) measures how a series is correlated with itself at different lags. "Correlation values, called correlation coefficients, can be calculated for each observation and different lag values. Once calculated, a plot can be created to help better understand how this relationship changes over the lag.



Partial Autocorrelation plot



. Data Preprocessing

- Applied Box-Cox transformation to normalize VWAP.
- Smoothed the series using moving averages.
- Conducted autocorrelation (ACF) and partial autocorrelation (PACF) analysis to examine lag dependencies.
- Tested for stationarity using the Augmented Dickey-Fuller (ADF) test.
  The **ADF test has a null hypothesis (H_0)**: the time series has a unit root, i.e., it is non-stationary.
  The **alternative hypothesis (H_1)**: the time series is stationary.
  Test statistic value: -2.694529
  A more negative value indicates stronger evidence against the null hypothesis (i.e., more likely the series is stationary).
  If the ADF statistic is less than the critical value at a chosen significance level (e.g., 1%, 5%, 10%), you reject H_0 → series is stationary.
- If it's higher than the critical value, you fail to reject H_0 → series is non-stationary.
- Performed time series decomposition to extract level, trend, and seasonality components.

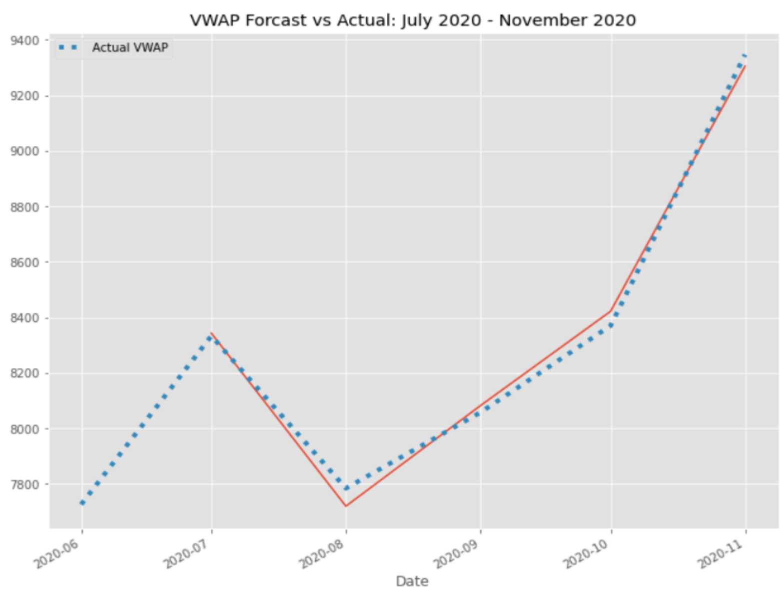
Preparing Model

Traditional Time Series Models: ARIMA, SARIMA, AutoRegressor, and ARIMAX were implemented to capture linear temporal dependencies in VWAP.

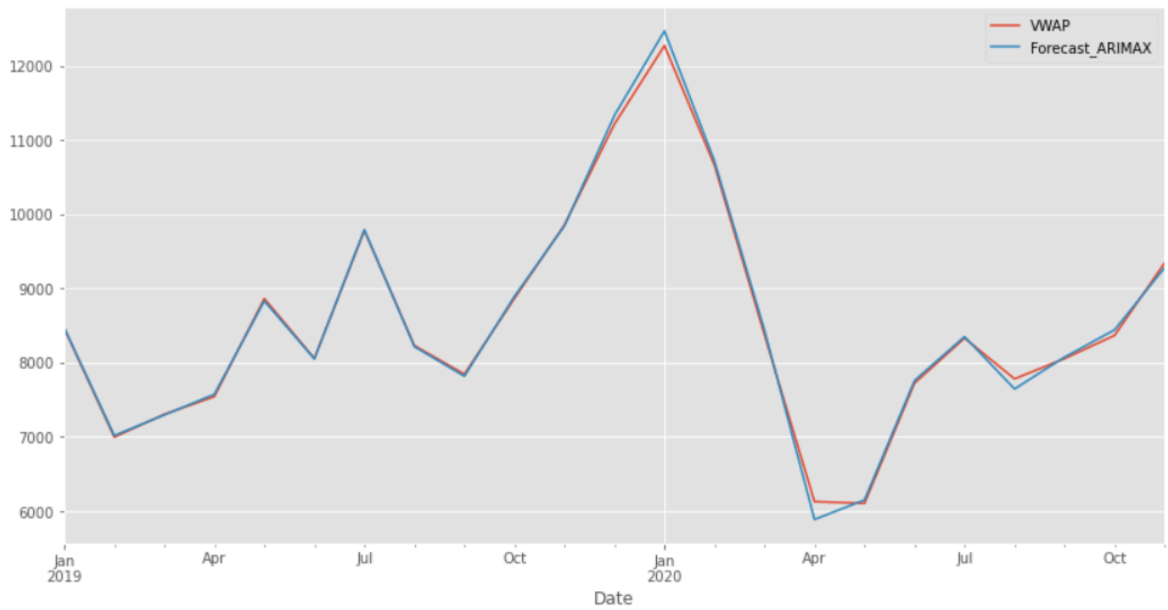
Deep Learning Models: LSTM and RNN were used to capture complex, non-linear patterns and long-term dependencies in the time series.

Comparison: Both traditional and deep learning approaches were trained and evaluated using RMSE and MAE to identify the best-performing forecasting model.

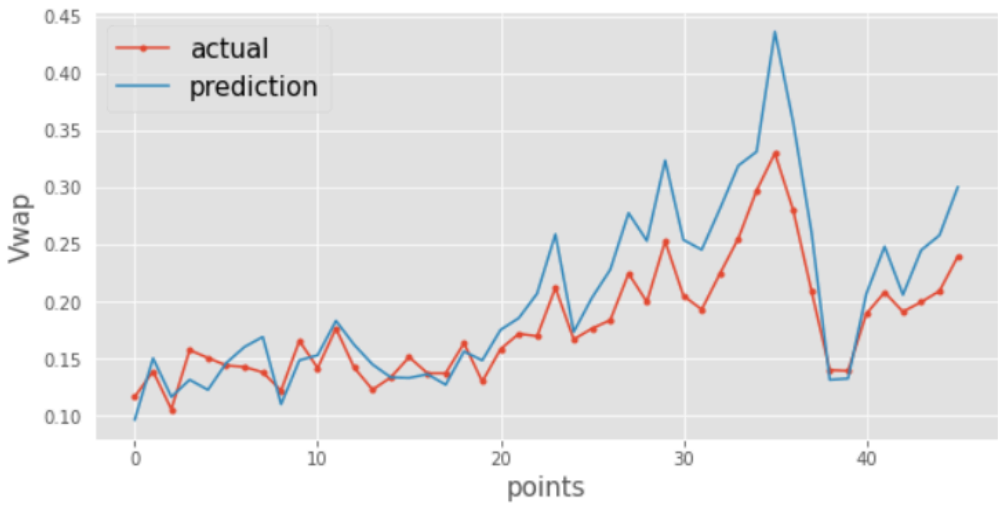
Forecasting from Autoregressor -



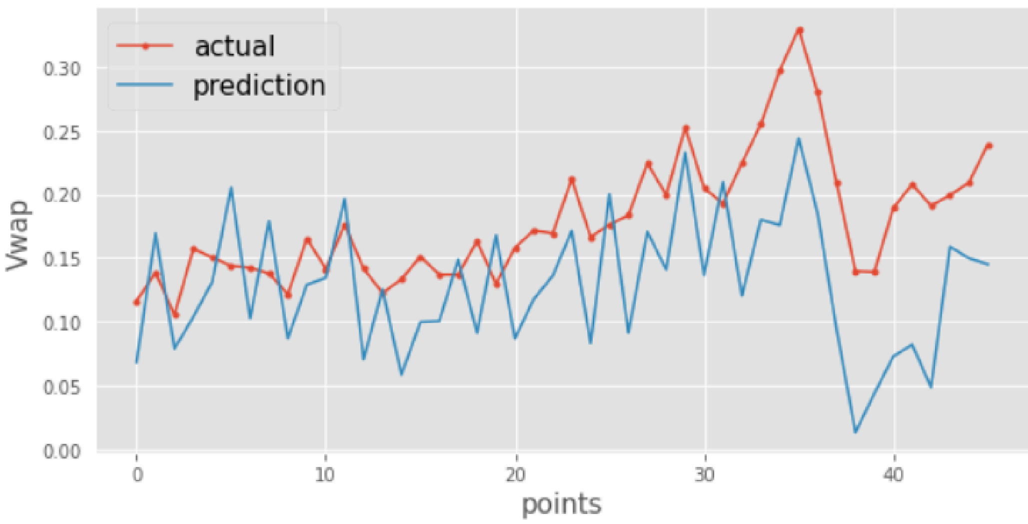
Forecasting from ARIMA -



Forecasting from LSTM



Forecasting from RNN-



Model Evaluation -

Statsmodels Autoregressors: rmse score - 32.30

pmdarima ARIMAX: rmse score - 82.27 mae score - 55.83

keras LSTM sequential model: rmse score - 0.046 mae score- 0.0370

keras RNN sequential model: rmse score - 0.046 mae score- 0.0624

Model	RMSE	MAE
Auto Regressor	32.30	–
Auto ARIMAX	82.27	55.83
LSTM	0.052	0.043
RNN	0.038	0.061

Insights:

- LSTM achieved best overall prediction accuracy, effectively capturing long-term dependencies and non-linear relationships.
- RNN showed slightly lower RMSE than LSTM but higher MAE, indicating less consistent predictions.
- Traditional models like ARIMA and ARIMAX had higher errors and were less capable of modeling complex temporal patterns.

Conclusion

- Both traditional and deep learning models were implemented and compared for VWAP forecasting.
- LSTM demonstrated superior performance, combining flexibility and memory control to capture temporal dynamics effectively.
- The workflow—from preprocessing, feature engineering, and modeling to evaluation—provides a comprehensive approach to time series forecasting in financial datasets.