Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

VWAP Forecasting for a Stock using Machine Learning

Harsh Joshi
Student, Computer Engineering
Dharmsinh Desai University, Nadiad
Gujarat, India

***Abstract* — The Indian National Stock exchange carries out a mean of 2,58,95,350 trades daily, amounting to a turnover of about 2000Cr, offering more than 7950 stocks to choose from. Stock market investment strategies are elaborate and rely on an assessment of vast amounts of information. In recent years, machine learning techniques have increasingly been examined to assess whether they can improve market forecasting when compared with other traditional approaches. Existing techniques like sentiment analysis or neural network techniques can be too narrow in their approach and can lead to erroneous predictions. A plethora of factors need to be considered before choosing a methodology and designing a model for forecasting stock's volume-weighted average price.**

Keywords: VWAP(Volume Weighted Average Price), RNN(Recurrent Neural Network), MA(Moving Average), LBGM (Light Boosting Gradient Machine), Facebook Prophet, Augmented Dickey-Fuller Test, LSTM (Long Short Term Memory), RMSE (Root-mean-squared Value), MEA (Mean Absolute Error), ARIMA (AutoRegressive Integrated Moving Average).

INTRODUCTION

As long as markets have existed, investors have been in constant search of new ways to acquire knowledge about the companies listed in the market to improve their investment returns. In the past, investors relied upon their personal experience to identify market patterns, but this is not feasible today due to the size of the markets and the speed at which trades are executed. The world's stock markets encompass enormous wealth. Time series analysis of this data can prove to be insightful. Time series analysis is a statistical technique that deals with trend analysis time-series data. The data is classified based on its inherent differences into the following three categories :

- **Time series data:** Time series data is a set of observations on the values that a variable takes at different times over a particular
- **Cross-sectional data:** It is the data of one or more variables that have been collected at the same particular point in time.
- **Pooled data:** A combination of cross-sectional data and time-series data.

The **objective** of this study is to assess various machine learning techniques available and figure out which one could provide a more robust prediction and design a model taking into account the myriad of factors influencing the market and based on this analysis, design a model that precisely forecasts the volume-weighted average price of a stock(TCS).

I. DATASET

The dataset used here is stock market data of the Nifty-50 index from NSE (National Stock Exchange), India over the last 20 years (2000 - 2019).

Feature	Description
Date	Date of the trade
Symbol	Symbol of the listed company
Series	Series of the equity(EQ, BE, BL, BT, IL)
Prev Close	The previous day closing price.
Prev open	Starting price at which a stock is traded in a day
High	The highest price of equity symbol in a day.
Low	The lowest price of the share in a day
VWAP	Volume weighted average price
Deliverable	Actual portion of total traded volume into Demat

The target value to predict here is the VWAP (Volume Weighted Average Price). Volume weighted average Price is a trading benchmark used by traders. Based on both, volume and price.It gives the average price at which the stock has traded at, throughout the day. It is an important parameter as it provides accurate insights into both the trend and value of the stock.

II. DATA ANALYSIS

FIg 2.1 TCS Stock Price

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Even though turnover and volume increased during the year 2020-2021, the prices dropped significantly. One possible explanation could be: Due to the pandemic, many investors took the advantage of dropping prices and bought stocks in bulk, perhaps looking forward to selling when industries regain momentum.

Exploring the missing values, trend, seasonality, correlation, and noise in the data:

1) Autocorrelation: Autocorrelation is a representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. Thus the following graph provides us a measure of the mathematical relationship between the present value and the past values.

2) Moving-Average: The moving average (MA) is a technical analysis tool. It smooths out price data by synthesizing a constantly updated average price. Abstractly it is a trend indicator which is an average of closing prices in a time frame. It can help identify a trading opportunity. Although TCS has great fundamentals, price/equity ratio, consolidation was observed in the last few months which is evident in the graph below:

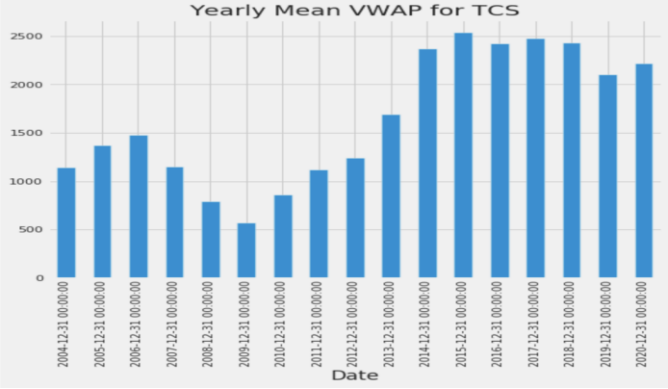


Fig 2.3 Yearly Mean VWAP or TCS

3) Stationarity Augmented Dickey-Fuller test: It helps identify stationarity in the dataset. A stationary time series is a series for which the statistical properties such as mean, variance, autocorrelation, etc remain constant over time.

- Null Hypothesis (H0): If failed to be rejected, it suggests that the time series has a unit root, implying that it is non-stationary and that It has some time-dependent structure.
- Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests that the time series does not have a unit root implying its stationarity i.e the fact that.The structure here is not time-dependent..
- Observed Values for TCS:
  - ADF Statistic : -1.7882201,
  - p-value: 0.386335
- Critical Values :
  - 1%:-3.432,
  - 5%:- 2.862,
  - 10%: -2.567

The test static value turns out to be -1.78. If random, such autocorrelations should be near zero for all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero. The more negative this statistic, the more inclined I was towards rejecting the null hypothesis (as that would imply that I have a stationary dataset).

Observation: The values are relatively not random but depend on the prior recorded values as ascertained by the augmented Dickey-Fuller test. The p-value turns out to be 0.386335 or 38.63%..Small p-value provides us with strong evidence to reject the null hypothesis, thus it is an evidence against the null hypothesis.

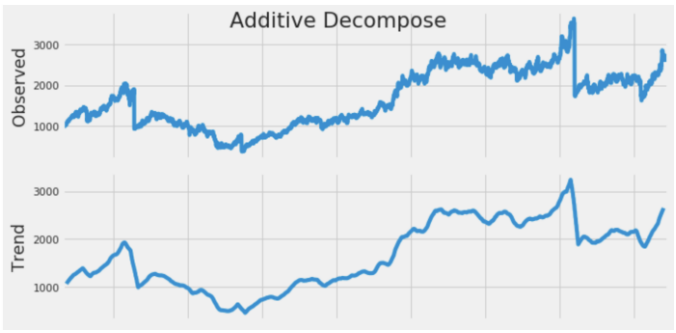
4) Missing values: Deliverable Volume and Deliverable percentage are the columns with missing values. 21% of trade data is missing, while only 10% of each of deliverable volume and the deliverable percentage is missing. The rows where deliverable volume is missing can be dropped.

III. MACHINE LEARNING ALGORITHMS EXPLORED

1) Facebook Prophet: Developed by Facebook, Prophet is an open-source time series model. It is based on decomposable (trend+seasonality+holidays) models. It provides us with the ability to make time-series predictions with good accuracy using simple intuitive parameters and also takes into consideration the impact of custom seasonality and holidays. It works best with time-series that have strong seasonal effects and several seasons of historical data which aligns with our needs as TCS is generally attributed with these qualities.

Time-series seasonal decomposition: We can decompose a time series into trend, seasonal, and remainder components. The time series can be decomposed as an additive or multiplicative combination of the base level, trend, and seasonal index and residual. The seasonal_decompose in the model is used to implement the decomposition.

The Root Mean Squared Error turned out to be **161.26780711** for the model whereas the Mean Absolute Error was found to be **109.83921808112.**



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

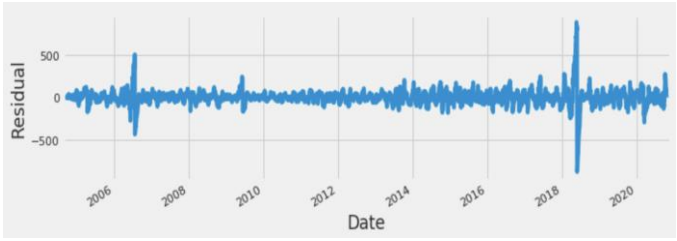


Fig3.1 Observed, Trend and Residual Additive decomposition

2) LightBGM: LightBGM, short for Light Gradient Boosting Machine, is an open-source distributed gradient boosting framework for machine Learning by Microsoft. It is based on decision trees to increase the efficiency of the model and reduce memory usage.

In our case, LightGBM performs terribly!. This exemplifies an important aspect of using boosting models for time series. Boosting models are constrained to predict within the range of target values appearing in the training data. The maximum price value in the training data is ~ 3100 and hence LGBM is unable to predict values beyond 3100.

The Root Mean Squared Error turned out to be **1233.324223074** for the model whereas the Mean Absolute Error was found to be **959.1349439727**.

3) AutoRegressive Integrated Moving Average: ARIMAmodel explains a given time series based on its past values, that is, its lags and the lagged forecast errors, so that equation can be used to forecast future values.ARIMA models require certain input parameters: p for the AR(p) part, q for the MA(q) part, and d for the I(d) part. Auto ARIMA is an automatic process by which these parameters can be chosen. When exogenous regressors are used with ARIMA it is commonly called ARIMAX.

The Root Mean Squared Error turned out to be **147.086385890** for the model whereas the Mean Absolute Error was found to be **104.0194218424.**

4) **Recurrent Neural Network**: RNN is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. As a result of this RNN exhibits temporal dynamic behavior. Derived from feedforward.RNNs can use their internal state (memory) to process variable-length sequences of inputs.

The Graphical Representation of RNN loss for TCS stock data can be observed below:

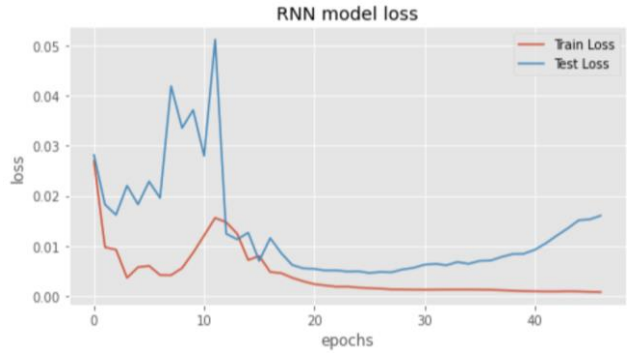


Fig3.2 RNN model Loss

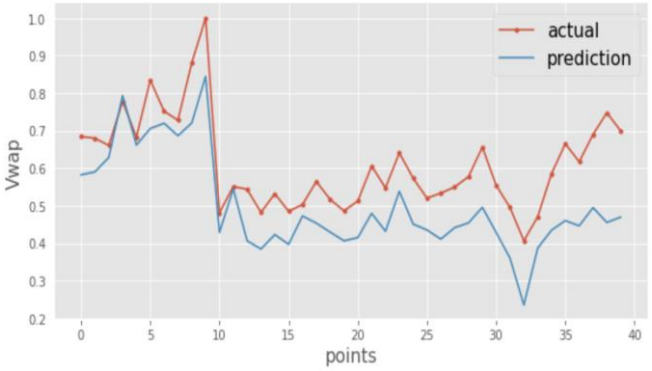


Fig3.3 Actual Vs Prediction

5) XGBoost: model takes the first fifty values as input and predicts the next value. For that, I had prepared the data with sixty previous values as 'X' and the current value as 'Y'.I took the last 7 months for Testing from 02-11-2019 to 31-06-2020 and remaining for the training 2960 days. Then, I tried to predict the next day, which is 01-07-2020.

Mean Absolute Error: 179.86173. Mean squared Error: 90328.18087. The model did not perform well. However, the train_test split has its drawbacks. Because this approach introduces bias as I was not using all of our observations for testing and also we're reducing the train data size. Cross-validation can be used to solve this. In Cross-validation, all the data is used for training and testing periodically. Thus there was scope to reduce the bias introduced by train_test_split. From different cross-validation methods, I used k-fold cross-validation.

In order to optimize the Hyperparameters, random Search was used with a target of optimizing their hyperparameters and thus improving their accuracy. Retraining using the best hyperparameter value and predicting stock value for the next day, using data from the last 50 days gave better results with Mean absolute error - 134.11.

CONCLUSION

The following table lists the RMSE and MAE values for the various models tested.

Model	RMSE	MAE
ARIMAX	147.086385890	104.0194218424
Facebook Prophet	161.267807113633	109.82921808112
LightBGM	1233.3242230741	959.1349439727
RNN	198.0424492431	112.525299191
XGBoost	179.86173	110.2348775

Applying the following traditional techniques do not make much sense because the fitting of time series models can be an ambitious undertaking :

- Box-Jenkins ARIMA models
- Box-Jenkins Multivariate Models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

- Holt-Winters Exponential Smoothing (single, double, triple) The user's application and preference will decide the selection of the appropriate technique.

LightGBM performs terribly! This is a very important aspect of using boosting models for time series. Remember that boosting models are constrained to predict within the range of target values appearing in the training data. The maximum price value in the training data is ~ 3500 and hence LGBM is unable to predict values beyond 3500.

They fail only in cases where the trend component is extremely strong and there are a wide variety of use cases where the trend is weak and the expected forecasts are within the values of the past. Stock prices are an example that generally has strong trend components, especially when measured over years. Here I tried to compare models by their RMSE and MAE value but the trade-off here is that RMSE accounts for a large number of outliers and accommodates them while fitting the model but it does not describe the average error alone and has a few more implications that are difficult to tease out and understand.

REFERENCES

[1] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for gradient boost classification. arXiv preprint arXiv:1801.06146, 2018.

[2] Shanker Iyer, Nikil Dandekar, and Kornl Csernai. quora dataset release, January 2017. URL https://www.quora.com/q/quoradata/ First-Quora-Dataset-Release-Question-Pairs.

[3] Ben Roshan, Augmented Dickey-Fuller - Time series analysis. https://www.kaggle.com/prashant111/arima-model-for-time-series-forecasting

[4] Vikas Singh, stock price prediction using xgboost, facebook prophet, Altair.

[5] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. arXiv preprint arXiv:1907.10529, 2019.

[6] Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning, 2012

[7] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in Translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems 30, pp. 6294–6305. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/ 7209-learned-in-translation-contextualized-word-vectors.pdf.