Modelling direction detection in selected stocks in

Indian BFSI sector

Anand Mohan  
*REVA Academy for Corporate Excellence(RACE)*  
*REVA University*Bangalore,India  
anandmohan.embedded@gmail.com

J.B Simha  
*REVA Academy for Corporate Excellence (RACE)*  
*REVA University*Bangalore, India  
jbsimha@gmail.com

ShinuAbhi  
*REVA Academy for Corporate Excellence (RACE)*  
*REVA University*Bangalore, India  
shinuabhi@reva.edu.in

*Abstract*— several research initiatives have been taken to predict stock market returns using historical data. During this capstone project, twenty-two years' price of the stock's daily close price is being utilized for direction detection. The objective of the project is to get the right stock and understand the data pattern using Exploratory Data Analysis and perform data preparation and then build the right models by using multiple Modelling techniques to predict whether the price will move up or move down. Closing prices is being utilized as 6 different feature variables for building the classification Model. The difference between the 7th and 8th-day Closing price is determined.0.7% difference,1% difference, and 1.5% difference are different classes of direction for which the rule is being set to determine either positive change, negative change, or no change. Similar process is again repeated for a range of consecutive days to be utilized as the feature variable increased to 10 days and 14 days. Then momentum, trend, volatility and volume indicators are being utilized as feature variables and different classification models are being built to determine upward direction detection. Random forest modelling has given the highest efficiency in direction detection. Logistic regression modelling done for percentage change in close price as 0.5% has given the highest efficiency for volume and momentum indicators whereas XG Boost Classifier provided the best prediction performance for trend and volatility indicators. The invaluable take away from the capstone is that various classification modelling techniques had been remarkably useful in direction detection for the stock under consideration.

Keywords— Direction detection, Stock Market, Technical Indicators, Classification Models, HDFC, KOTAK, SBI

# Introduction

Live validations are still becoming a grim prospect, because of several things like value variations, quiet news, and existing noise (Shah et al., 2019).

A number of Machine-Learning associated techniques are developed which have created the potential to predict the market to an extent (Sonkiya et al., 2021).

For the transaction of shares via a broker, there is mostly a fee paid to the broker for each buy and sale which will almost eat up the gains (Huang et al., 2021).

The requirement is to overcome the ambiguities of Fundamental and technical evaluation, and advanced development in the modelling strategies has pushed several researchers to check new strategies for stock value forecasting (Rouf et al., 2021).

In the next section, some of the available literature will be scanned which would throw light on various related aspects of Machine-Learning methods and other methodologies, and also study and research other related issues which would help assist better in direction detection in Stock Market.

# LITERATURE REVIEW

## Algorithmic trading

Ultrafast algorithms improve traders’ ability to seize opportunities long before any human would be able to do the same(Hansen, 2020).

Regulators have restrained algorithmic commerce, following accusations of market manipulation(Mukerji et al., 2019).

## Fundamental analysisof the stock market

Fundamental analysis helps to identify and implement short positions by selling the shares of companies showing downtrends and then covering these positions by buying back the shares of these companies when they start showing upward trends (Elbialy, 2019).

## Technical indicators for the stock market

The expectation of various crypto currencies like Bitcoin, Ethereum, Litecoin, and Ripple digital currency value in examination with the anticipated price by the volatility regression model and trend indicators gave pretty higher returns for the entire month (Dahham & Ibrahim, 2020).

Momentum-based Trading commerce is amongst proved investment strategies across major stock markets (Mohapatra & Misra, 2020).

## Supervised and Unsupervised learnings

Some literature has used both supervised and unsupervised machine learning techniques for securities market predictive modelling (Alhomadi, 2021).

## Principal Component analysis

The central plan of PCA is to spot correlations and patterns in a dataset with high dimensionality and scale back it to a considerably lower dimension without losing any important info (Dar, 2021).

## Logistic Regression

LR is used instead of linear regression in situations where the target variable is not numeric, but a nominal or an ordinal variable (Al-Bairmani & Ismael, 2021).

## Decision Tree

In Decision Tree, the model becomes more complex as the size of the datasets increases. This is being handled using more advanced algorithms in Decision Tree for classification and regression problems (Jena & Dehuri, 2020).

## Random Forest

RF is quite flexible to non-linearity in the dataset and is the most appropriate ensemble learning algorithm for medium-sized to very large-sized datasets (Schonlau & Zou, 2020).

## K-Nearest Neighbours

K-Nearest Neighbors is the most popular statistical technique utilized in pattern identification over the last four decades (Wang, 2019).

## Extreme Gradient Boosting

XGBoost is extensively recognized as an extremely useful ensemble learning algorithm. However, its performance needs more improvements ideally in scenarios where the dataset is imbalanced (Zhang et al., 2022).

## confusion matrix for Classification Models

The confusion matrix evaluates numerous performance metrics which include accuracy, precision, and recall (Markoulidakis et al., 2021).

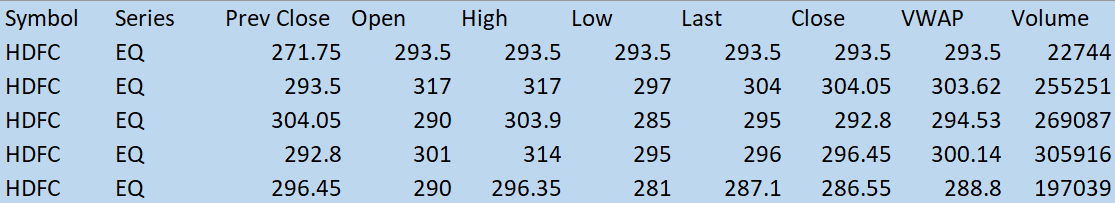
# METHODOLOGY

Initially Fundamental and Technical analysis of HDFC stock is performed to demonstrate why the HDFC stock dataset has been used for this project. Data understanding explains the different columns used in the HDFC dataset. Data preparation explains that Handling Missing values, Features Addition and Data Scaling using MinMax Scaler were the steps used for processing the dataset before being used for Modelling. Hypothesis testing, Classification Models, ARIMA Models, and different Regression Models were used in the Data Modelling phase. The data evaluation phase examines the results of different Modelling techniques which were used in the Data Modelling phase. Deployment speaks about developing a front-end API for the deployment Dashboard.

## Data Collection

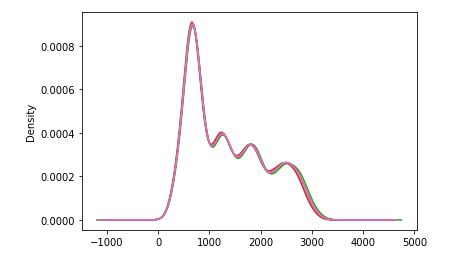
Daily Trading Data of HDFC company from the year 2000 to 2021 is being used for this study. This study uses NSE Data.

The symbol column tells us the corporate symbol mentioned for the stock. The opening price is the first trade worth that was recorded throughout the day’s trading. The high is the highest worth at that a stock is listed during a period. The low is the lowest worth of the period. The previous closing is going to be a consecutive session's opening price. The last price is the one at which the foremost recent transaction happens. The close is the last commerce worth recording once the market is closed on the day. The volume-weighted average worth (VWAP) is a trading benchmark based on both volume and worth. Trading Volume shows the number of shares listed for the day, listed in lots of 100 quantities of shares.



1. Top 5 rows of HDFC stock dataset

## Data Exploration



1. Data Distribution plot of Feature variables and Close price for HDFC Stock

The mean value is greater than the median value of most of the feature variables which is represented by the 50th percentile meaning Data has a positively skewed distribution. There is notably a large difference between the 75th percentile and max values of most of the feature variables. These observations suggest that there are extreme values-Outliers in our data set.

## Data Pre-processing

The HDFC data which is taken from NSE comes with a lot of limitations and that has to be processed which includes the following steps:

Handling Missing values: Three of the features’ trades, ‘Deliverable Volume’, and’% Deliverable had quite one hundred periods of missing values therefore those columns need to be dropped as they are having several missing values.

Features Addition: Additionally, computed variables were added to the dataset that for sure would influence stock returns. These are moving averages for rolling periods of seven days,13 days,20 days,100 days, and two hundred days. conjointly enclosed were EMA for seven days,13 days,20 days,100 days, and two hundred days. one day's previous lag values of volume are also added in the concert of the input feature.

Data Scaling: Minmax Scaler is the data scaling approach that is being used. Here, the minimum of features is created up to zero, and the most of features are up to one. MinMax Scaler shrinks the data inside the given range, from zero to one.

## Data Modeling

A rule-based model is being developed to do hypothesis testing to determine whether the chosen stock's price is crossing any of the following moving averages: the 7-day, 13-day, 20-day, 100-day, and 200-day moving averages. It will be a purchase decision if the projection indicates that the value will be higher than various Moving Averages. Exponential Time series Models are used to create the same five hypothesis testing models. After that, five further ARIMA-based time series models are created to support the buy or sell recommendation for every stock.

Various Classification models namely AutoKeras Classification Model, K-neighbours Classifier Model, and Logistic Regression Classification Model deployed and their prediction accuracy is being compared with SMA Models, EMA Models, and ARIMA Models.

Further ahead various Regression Models including both Machine Learning and Deep learning techniques are deployed and Metrics namely MAE and MAPE are deployed to estimate the quality of the predictions on the close price of the HDFC share. These Regression Models are the OLS-Linear Regression Model, Lasso Regression Model, Lasso regression Model Using Cross Validation, The KNN Algorithm, Decision Tree Algorithm, GridSearchCV Algorithm with Hyperparameter Tuning, Random Forest Regression Model, XGBoost ML Model, Using PCA with LSTM, Using PCA with LSTM with Moving Average variables (Feature Engineering), LSTM Neural Network Model, Regression Model using AutoKeras.

# FINDINGS/DISCUSSION

The Data Evaluation phase is the results of the Data Modelling phase and discusses the Metrics utilized to determine the extent of successes achieved from the different Modelling Algorithms employed on the Target Variable.

## SMA EMA T Test Metrics

The hypothesis testing rule’s accuracy is repeatedly verified. The T-test is employed to perform hypothesis testing for SMA of 7 days.13days, and 20 days and EMA with 7,13 days, and 20 days spans are employed to recreate the various models based on T-test Hypothesis Testing.

| Serial Numbers | Total | True Count | False Count | Efficiency |
| --- | --- | --- | --- | --- |
| **SMA7** | **5297** | **4114** | **1183** | **77.67** |
| *SMA13* | *5291* | *3474* | *1817* | *65.66* |
| *SMA20* | *5284* | *3217* | *2067* | *60.88* |
| *EMA7* | *5297* | *4077* | *1220* | *76.97* |
| *EMA13* | *5291* | *3486* | *1805* | *65.89* |
| *EMA20* | *5284* | *3236* | *2048* | *61.24* |

Table1. Leader Board-comparison of Metrics for SMA and EMA variables as per T Test based on Hypothesis Testing

From Table 1, It can be observed that T-test Hypothesis testing done for 7-days SMA has given the highest efficiency in correctly predicting the upward or downward trend closely followed by 7-days EMA. However, prediction efficiency is the least for 20-day SMA and 20-days EMA.

## SMA EMA Z Test Metrics

#### The hypothesis testing rule's accuracy is repeatedly verified. Z-test is employed to perform hypothesis testing because the sample size for testing is more than 30 samples. SMA of 100,200 days and EMA with 100 days and 200 days spans are employed to recreate the various models.

| Serial Numbers | Total | True Count | False Count | Efficiency |
| --- | --- | --- | --- | --- |
| SMA100 | 5204 | 2798 | 2406 | 53.77 |
| *SMA200* | *5104* | *2754* | *2350* | *53.96* |
| *EMA100* | *5204* | *2829* | *2375* | *54.36* |
| *EMA200* | *5104* | *2779* | *2325* | *54.55* |

Table2. Leader Board-comparison of Metrics for SMA and EMA variables as per Z Test based on Hypothesis Testing

From Table 2, It can be observed that Z-test Hypothesis testing done for a rolling 100-day moving average and 200-day moving average has given lesser efficiency in correctly predicting the upward or downward trend compared to the prediction done with Hypothesis testing done on smaller samples using T-test Hypothesis testing. Similar inferences can be drawn for EMA with 100 days and 200 days span as well.

## Classification Model Metrics

Auto Keras Classification Model, KNN Classification Model, and Logistic Regression Classification Modelling techniques are deployed to predict the direction of the close price.

| Serial Numbers | Total | True Count | False Count | Efficiency |
| --- | --- | --- | --- | --- |
| **Auto Keras** | **1061** | **901** | **160** | **84.92** |
| *KNN* | *1061* | *786* | *267* | *74.08* |
| ***LR*** | ***1061*** | ***956*** | ***97*** | ***90.10*** |

Table3. Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by different Classification Models

From Table 3, It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price. The highest Accuracy in predicting the direction by Hypothesis Testing using SMA and EMA was near about 77%. Hence, it can be safely concluded that Deep Learning models and Machine Learning Models were able to provide better outputs compared to Statistical methods of Hypothesis Testing.

## ARIMA Models Metrics

In all results of the ADF test for ARIMA Modelling on the dataset for HDFC stock, the p-value obtained was bigger than 0.05 thus the null hypothesis is not rejected, and concluded that the statistic for Dataset under consideration is non-stationary. Also, MAE, MSE, RMSE, Median Absolute Error, and MAPE are far too high in the case of all Auto ARIMA Modelling. Hence, it can be concluded that the dataset under consideration was not suitable for Time series Modelling using the ARIMA Modelling algorithm.

## Regression Models Metrics

OLS-Linear Regression Model,Lasso Regression Model,Lasso regression Model Using Cross-Validation and KNN regression Models are deployed to predict the close price.

| Serial Numbers | MAE | MSE | RMSE | Median Absolute  Error | MAPE |
| --- | --- | --- | --- | --- | --- |
| **OLS** | **2.03** | **11.83** | **3.44** | **1.14** | **0.23** |
| *LASSO* | *7.56* | *132.63* | *11.52* | *4.67* | *0.85* |
| *LASSOCV* | *7.55* | *132.59* | *11.51* | *4.66* | *0.85* |
| *KNN* | *5.42* | *132.08* | *11.49* | *3.16* | *0.59* |

Table4. Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by the First set of Regression Models

From Table 4, It can be observed that MAE and MAPE were satisfactory for the OLS-Linear Regression Model. However, other Regression Models were not able to provide MAPE within the acceptable range.

All the models are now combined and below is the description for the final results.

## Classification Metrics Comparison

| Serial Numbers | EFFICIENCY>67% |
| --- | --- |
| **SMA-7 samples** | **YES-77.67** |
| *SMA-13 samples* | *NO-65.66* |
| *SMA-20 samples* | *NO-60.88* |
| EMA-7 samples | *YES-76.97* |
| *EMA-13 samples* | *NO-65.89* |
| *EMA-20 samples* | *NO-61.24* |
| SMA-100samples | *NO-53.77* |
| *SMA-200 samples* | *NO-53.96* |
| EMA-100 samples | *NO-54.36* |
| *EMA-200 samples* | *NO-54.45* |
| ***Auto Keras*** | ***YES-84.92*** |
| *KNN* | *YES-74.08* |
| ***LR*** | ***YES-90.10*** |

Table5. Leader Board-comparison of Metrics for Classification Models

From Table 5, It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price. The highest Accuracy in predicting the direction by Hypothesis Testing using SMA and EMA was near about 77%. other Hypothesis testing using T-test and Z-test statistical algorithms were not satisfactory in able to predict the direction of the close price of the HDFC share.

## Regression Metrics Comparison

| Serial Numbers | MAE<=5 | MAPE<=0.33 |
| --- | --- | --- |
| **OLS** | **YES-2.034** | **YES-0.23** |
| LASSO | NO-7.555 | NO-0.85 |
| *LASSOCV* | *NO-7.55* | *NO-0.85* |
| *KNN* | *NO-5.423* | *NO-0.59* |
| DT | *YES-3.26* | *NO-0.38* |
| *GridSearchCV* | *YES-3.218* | *NO-0.38* |
| ***RF*** | ***YES-2.45*** | ***YES-0.29*** |
| XG Boost | *YES-3.25* | *NO-0.37* |
| ***LSTM using PCA*** | ***YES-4.366*** | ***YES-0.33*** |
| LSTM using PCA with moving average variable | *NO-7.75* | *YES-0.33* |
| *LSTM* | *NO-9.71* | *YES-0.33* |
| ***Auto Keras*** | ***YES-2.59*** | ***YES-0.27*** |

Table6. Leader Board-comparison of Metrics for Classification Models

From Table 6, It can be observed that the OLS-Linear Regression Model, Random Forest Regression Model, Using PCA with LSTM, and Regression Model using AutoKeras provide MAE<=5 and MAPE<=0.33. Hence these Regression Models were most successful in predicting the close value of the stock price. XGBoost ML Model, Decision Tree Algorithm, GridSearchCV Algorithm with Hyper-parameter Tuning provided good MAE but were slightly higher with MAPE.

# CONCLUSION/IMPLICATIONS

The hypothesis testing rule's percentage accuracy was repeatedly verified using five SMA Models. EMA was used to recreate the five other different models created using SMA. T-test was used to perform hypothesis testing if the sample size for testing was lesser than 30 samples. Z-Test was used to validate null and alternate hypothesis testing for samples larger than 30.ARIMA Time series modelling was used to create an additional five different models. The construction of all 15 models, was used to forecast day trading in the stock market. Prediction accuracy was then compared with Classification Model Algorithms. When the majority of the various models or all of them move in the same direction, a choice on whether to purchase or sell the stock must be made.

This paper then solely focuses on predicting the close price of the HDFC stock using Regression algorithms deploying both Machine Learning and Deep Learning Techniques. What works in the Indian stock market must be proven with evidence. Any stock on the stock market can utilize the same procedure to forecast buy or sell choices, which is helpful.

# RECOMMENDATIONS

It is assumed that returns are more or less constant over time. However, the assumption that the returns are constant over time is restrictive, and not true. Returns are highly dependent on time. In the future, it can be shown how to define Bullish and Bearish regimes using modern machine learning techniques. The Sentiment Analysis Approach may also need to be explored using Text Analytics for predicting stock market returns. In the Future, there is a deployment Dashboard proposed. An intelligent Automated system for Options Trading would be also the next step forward.

##### References

1. Alhomadi, A. (2021). Forecasting stock market prices : A machine learning approach. *Digital Commons*, *11*(2), 16–36.
2. Huang, Y., Capretz, L. F., & Ho, D. (2021). Machine Learning for Stock Prediction Based on Fundamental Analysis. *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*, *5*. https://doi.org/10.1109/SSCI50451.2021.9660134
3. Jierula, A., Wang, S., & Oh, T. (2021). applied sciences Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data.
4. López del Val, J. A., & Alonso Pérez de Agreda, J. P. (1993). Principal components analysis. *Atencion Primaria / Sociedad Española de Medicina de Familia y Comunitaria*, *12*(6), 333–338. https://doi.org/10.5455/ijlr.20170415115235
5. Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics (Switzerland)*, *10*(21). https://doi.org/10.3390/electronics10212717
6. Series, I. (2021). Machine Learning Algorithms and Applications. In *Machine Learning Algorithms and Applications* (Vol. 7). https://doi.org/10.1002/9781119769262
7. Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(3). https://doi.org/10.3390/ijfs7020026
8. Sonkiya, P., Bajpai, V., & Bansal, A. (2021). *Stock price prediction using BERT and GAN*. *6*. http://arxiv.org/abs/2107.09055
9. Vreeken, J., & Yamanishi, K. (2019). Proceedings of the 25th {ACM} {SIGKDD} International Conference on Knowledge Discovery & Data Mining, {KDD} 2019, Anchorage, AK, USA, August 4-8, 2019. 1946–1956. https://doi.org/10.1145/3292500
10. Сороко, Н. В. (2017). Масові Відкриті Європейські Он-Лайн Курси Для Вчителів (2017 Р.). Інформаційний Бюлетень№ 1. 801, 1–23.
11. Elbialy, B. A. (2019). The Effect of Using Technical and Fundamental Analysis on the Effectiveness of Investment Decisions of Traders on the Egyptian Stock Exchange. International Journal of Applied Engineering Research, 14(24), 4492–4501. <http://www.ripublication.com>
12. Faijareon, C., & Sornil, O. (2019). Evolving and combining technical indicators to generate trading strategies. Journal of Physics: Conference Series, 1195(1). <https://doi.org/10.1088/1742-6596/1195/1/012010>
13. Thanekar, G. S., & Shaikh, Z. S. (2021). Analysis and Evaluation of Technical Indicators for Prediction of Stock Market. International Journal of Engineering Research & Technology (IJERT), 10(May), 341–344.