Stock Price Prediction using Machine Learning

Authors,

[1]Akshay Bhimani, [2]Harsh Agola, [3]Darshak Chavda, [4]Hardik Modhvadia
[1,2,3,4]B.Tech - Information and Communication Technology, Ahmedabad University
Ahmedabad, India

email: {akshay.b, harsh.a, darshak.c, hardik.m}@ahduni.edu.in

***Abstract* - Stock market prediction or forecasting is a serious challenge for corporate brokers stakeholders and investors. The stock of each company fluctuates in a random manner and becomes hard and risky for investors to invest. In this project, we apply knowledge of machine learning techniques to predict the stock prices of a particular company. The stock market is so volatile and investors don't want to fall at risk. So an approach with adequate expertise is designed to help investors to ascertain veiled patterns from the historic data that have the feasible predictive ability in their investment decisions. The historical data has a significant role in, helping the investing people to get an overview of the market behavior during the past decade.**

Key Words - Machine Learning, Data set, ARIMA, Stock market, Prediction. Share Price, Business, Statistics, Investments

Introduction

A **stock market**, **equity market**, or **share market** is the aggregation of buyers and sellers of stocks (also called shares), which represent ownership claims on businesses. Major companies like Reliance, Jio, Tata consultancy services are IPO (Initial Public Offering) listed companies. For a company to share their partnership publicly they have to be IPO recognized and listed. A stock is a general term used to describe the ownership certificates of any company. Our goal in this project is to forecast or predict the value of the stock price of a particular IPO registered company. In this project, we are going to predict the stock price of Tata Consultancy Services (TCS) [1]. In this project, we are going to use machine learning to predict stock prices. Stock price prediction is one of the most difficult tasks and using machine learning algorithms to predict results in high accuracy and speed.

In particular, we are going to use a Linear regression model for the prediction of the stock value. Linear regression is a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). In our case, the dataset we acquired for TCS stocks from the year 2010 to 2021 suggests that linear regression can provide good accuracy in the prediction of stock price. Further, we have found that linear regression can not be suitable for this type of problem because it predicts close value only by looking at the open value and that can not be the case here. So we chose the ARIMA (Autoregressive Integrated Moving Average) model which is a subset of linear regression which attempts to use past observations of the target variable to forecast its future values. Here we have used the ARIMA model for short-term stock price prediction. It helps investors to make decisions for short-term stock investment.

Literature Survey

[1]The author has used linear regression, Polynomial, and RBF and compared its confidence values and it is found that Linear regression has more confidence value than the other two. Open, High, and low prices are input independent variables and close price is target-dependent variables.

[2]In this research paper, the author used supervised learning concepts like classification and regression. They have used KNN and linear regression to forecast the daily price of the stock and they have found that KNN-Algorithm shows 63% of accuracy, and Linear regression shows 98% of accuracy.

[3]In this research paper, the authors have used the tool Eviews software version 5 for the implementation of the ARIMA model. They have used two datasets for short-term stock price prediction, New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE). In the dataset, there are four elements, open price, close price, high price, and low price. Here, the closing price is selected to represent the price of the index to be predicted. The following criteria are used in this paper to define the best ARIMA model for each stock index: Relatively small of BIC (Bayesian or Schwarz Information Criterion), Relatively small standard error of the regression (S.E. of the regression), Relatively high of adjusted R2.

[4]The difficulty in deciding the order of an ARIMA model is discussed in this paper. The possibility of removing this difficulty by using the MAICE (minimum AIC estimation) procedure, which selects a model by using Akaike's Information Criterion (AIC), is checked with the numerical examples treated in the book by Box and Jenkins. The advantage of the B-J procedure lies in its use of the ARIMA model which is one of the most versatile models of time series. It was observed that the models adopted by the MAICE procedure from a wider set of ARIMA models also passed the B-J's diagnostic checking procedure, provided the checking procedure can be applied. Apart from that, there is one difficulty in dealing with this family of models that was the lack of an effective, objective criterion for selecting an optimal member. This paper shows that the MAICE procedure overcomes many of the difficulties of the order determination of ARIMA models experienced by users of the B-J procedure and provides a basis for the development of identification of higher-order ARIMA models.

IMPLEMENTATION

Firstly, we explored various datasets online and found a suitable dataset for the project. The Stock Dataset used here is Tata Consultancy Services. Published stock data obtained from Yahoo Finance are used with the stock price predictive model developed. After that, we have performed EDA and determined the correlation between the variables/parameters.

We also found that the stock market prices are time series in nature. Therefore, any suitable timed series model can be used to forecast the price of the stock. As for this project, we have used the ARIMA model.

It is also referred to as the Box-Jenkins methodology composed of a set of activities for identifying, estimating, and diagnosing ARIMA models with time-series data. The model is the most prominent method in financial forecasting. ARIMA models have shown the efficient capability to generate short-term forecasts. In the ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows[4]:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_0 + \phi_2 Y_{t-2} + \ldots + \phi_0 + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_p \varepsilon_{t-} \quad (1)$$

where Y_t is the actual value and ε_t is the random error at t, ϕ_i and θ_j are the coefficients, p and q are integers that are often referred to as autoregressive and moving average, respectively.[4]

To work with time-series data and to use the ARIMA model the data must be stationary and for that, we first checked the trends of the graph for the close value of stock price by days, months, quarters, and years. To check for stationery we have used the Dickey-Fuller test and calculated the moving average to conclude that the data is not stationary.

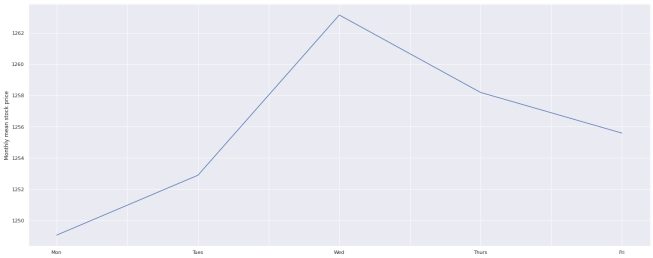


fig: average price on each day of week

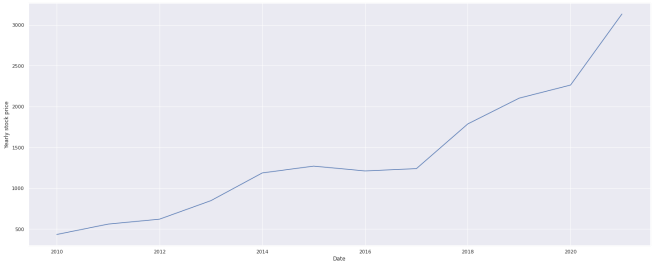


fig: stock price on each day

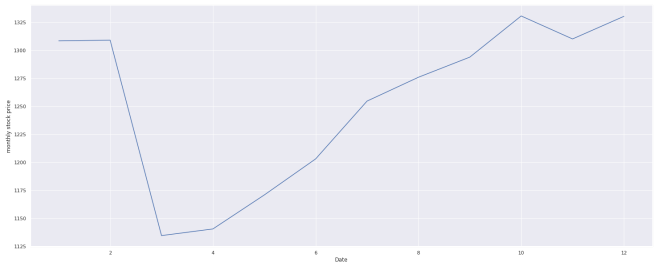


fig: stock price on each month

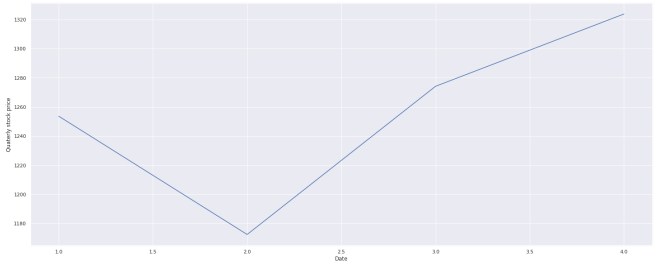


fig: stock price on each quarter

To make the data stationary we performed the following steps. We first do the logarithmic transformation of the close values of the dataset. We then took the rolling mean of the previously calculated log values. Then we took the difference between the former and latter to determine whether it is stationarity, we performed the dickey fuller test. We also calculated the exponentially weighted average on the logarithmic values. We calculated the difference between log values with log(t-5), where t is the date.

We then took seasonal decomposition to find the trend, seasonality, and residual of the calculated data. We then performed the Dickey-Fuller test on the residue value and concluded that the data has become stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                -1.847388e+01
p-value                        2.137967e-30
#Lags Used                     2.300000e+01
Number of Observations Used    2.723000e+03
Critical Value (1%)           -3.432754e+00
Critical Value (5%)           -2.862602e+00
Critical Value (10%)          -2.567335e+00
dtype: float64
```

After the determination of the stationarity, we found the p,q, and d values. Here p is autoregressive, q is moving average and d is the order of differentiation. To calculate these values we plotted ACF (Autocorrelation function) and PACF (Partial Autocorrelation function) on the residual data got from the previous step. The graph shows the values of the p and q as zeros.

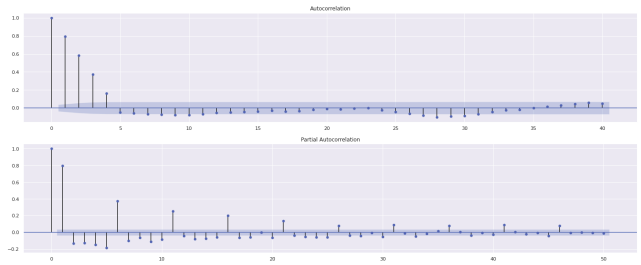


fig: Autocorrelation and Partial autocorrelation

We still checked the model with different values to get the best order possible for the model. To check for the best order we have trained the model several times and check for the ACI and BCI values. We found that the order of (0,1,0) was the best for the model and data we have used. To make sure we cross-checked it with the auto ARIMA function which helps to determine the best order.

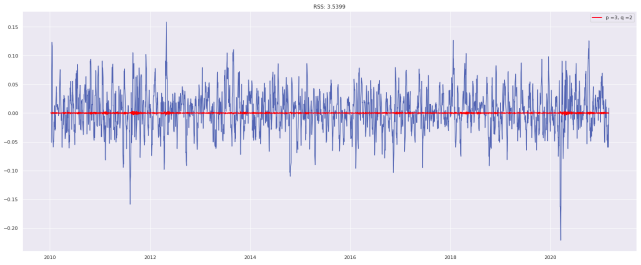
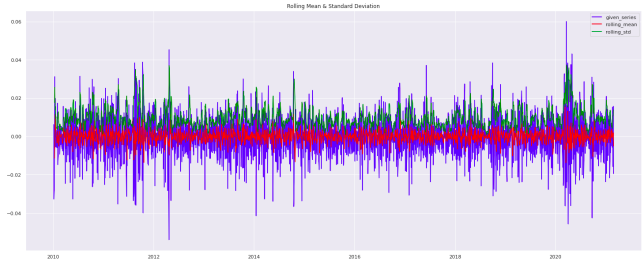


fig :

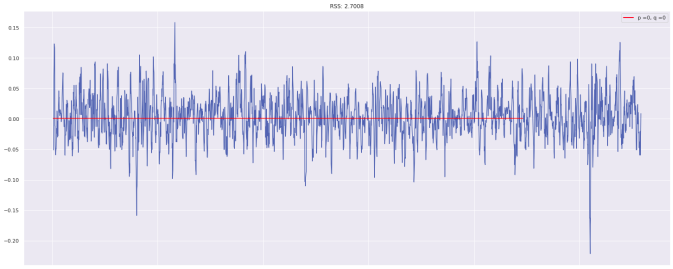


The training of the model was done by using the logarithmic values taken earlier and with the order (0,1,0). The forecasting was done with the inbuilt method of the stats model and Pmdarima library.

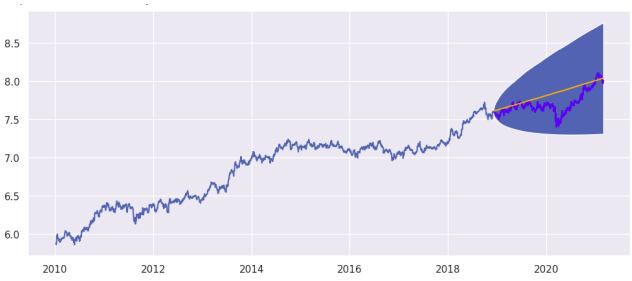
RESULTS

From our testing with different values of AR and MA we have found that p = 0 and q = 0 gives us best results

from dicky Fuller Test and also we have found that it has least RSS value.



Now we have performed prediction on final model and we have found that it gives good prediction for short term and fails if there is sudden change due to unexpected situations.




```
MSE: 0.03
MAE: 0.12
RMSE: 0.16
MAPE: 0.02
```

Here we found that MSE , MAE , RMSE and MAPE are very low and hence its good prediction.

Conclusion

This project presents the process of building an ARIMA model for stock price prediction for the TCS stock dataset. The experimental results obtained with the best order for the ARIMA model demonstrated the potential of ARIMA models to predict stock prices accurately for a short span of the time period. This can help in the investing and analysis for the investors to make a profitable investment. With the results obtained ARIMA models can compete reasonably well with emerging forecasting techniques in short-term prediction.

References

[1]"TATA CONSULTANCY S (TCS.NS) stock historical prices & data – Yahoo Finance," Yahoo! Finance, 18-Mar-2021. https://in.finance.yahoo.com/quote/TCS.NS/history?p=TCS.NS

[2] Y. Ng, "Machine Learning Techniques Applied to Stock Price Prediction," *Medium*, 03-Oct-2019. https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001.

[3] Poornima S P, Priyanka C N, Reshma P, Suraj Kr Jaiswal, and SurendraBabu K N, "Stock Price Prediction using KNN and Linear Regression". 2019 International Journal of Engineering and Advanced Technology (IJEAT), Volume-8, Issue-5S. [Online]. Available: International Journal of Soft Computing and Engineering (ijeat.org)

[4] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.

[5] T. Ozaki, "On the Order Determination of ARIMA Models," *Applied Statistics*, vol. 26, no. 3, p. 290, 1977.

[6] "Forecasting: Principles and Practice (2nd ed)," *8.6 Estimation and order selection*. [Online]. Available: https://otexts.com/fpp2/arima-estimation.html#arima-estimation. [Accessed: 11-Apr-2021].