Stock Market Prediction using ARIMA Model

Anoushka Dey

July 20, 2022

Introduction

This project uses the Auto-Regressive Integrated Moving Average (ARIMA) model to analyse and predict the stock prices of two stocks over a period of 6 months and 1 year. The initial weeks of the SoC were spent in installing and learning the basics of python and its libraries such as numpy, matplotlib, seaborn and pandas. After this, our mentors introduced us to the concept of exploratory data analysis through videos and other resources. Following this, we were also asked to try out some of the EDA on a stock using data from the yahoo finance API yfinance.
The time series ARIMA model uses historical data to analyse and predict stock prices using machine learning.
The project follows a certain workflow:

- Data Collection
- Formatting the data
- Plotting the ACF and the PACF
- Stationarizing the time series
- Building the ARIMA model by finding the order
- Making the prediction

The Workflow

I have used the ICICI and Reliance stocks for this project. The following sections will go through the process.

Data Collection and Formatting

First all the relevant python libraries were imported. These include numpy, seaborn, pandas, matplotlib, statsmodels, pmdarima and finally yfinance. The ICICI and Reliance databases were established using the *Ticker* function and accordingly adjusted to a timeline of 6 months or 1 year using the *history* function. The NA entries were subsequently dropped and the cleaned datasets were displayed. The *Date* and *Close* entries were group together to form the icici_6g, icici_1yg, rel_6g and rel_1yg datasets.

Plotting the ACF and the PACF

The data is said to be stationary if the mean, variance and autocorrelation do not show any change with respect to time. In short, it should not show any trends and should have a constant variance and autocorrelation. The Autocorrelation function (ACF) and the Partial autocorrelation function (PACF) graphs of all the four datasets were plotted. The Dickey-Fuller Test, although not shown, was performed and the conclusion drawn was that the four datasets were not stationary.

Stationarizing the time series

The four datasets had to be stationary for the ARIMA model to be applied on them. Hence the method of differencing was used. The number of differences done contributes to the integrated difference of the ARIMA model. After the differencing is performed, graphs are plotted to confirm stationarity of the datasets. From the graphs it is evident that the intervals are regular and that the stationarity of the datasets was established using the differencing method.

Building the ARIMA Model by finding the order

Next, using the *auto_arima* function, the values of p, q and d were found out for all the four datasets. The values found out are the best and they help optimize the ARIMA model of the datasets for efficient and correct forecasting. The (0,1,0) order was found to be most suitable for all the four datasets when the *auto_arima* function was applied to each of them. This would come in handy for building up individual models of the four datasets under consideration.

Let us consider the icici_6g dataset first. The *model* object is created using the ARIMA function by specifying the values and the order to be considered. Next, *model_fit* is established. This is subsequently followed by a joint plot of both the actual and the forecasted values of the stock prices of the ICICI bank for a period of 6 months. Here, I have used the *close* prices of the stock. It is evident that the forecasted and the actual values indeed do overlap to quite a large extent.
The same method has been applied to the icici_1yg, rel_6g and the rel_1yg datasets and the corresponding graphs of the prices have been plotted.

Making the Prediction

For each individual dataset, *model_fit* pertaining to itself has been used along with the *predict* function to predict the stock prices of the follwoing 30 days. The start and end indices have been specified and the entire set of stock prices has also been displayed.

Timeline

The learning phase took up most of the time allotted for the project. In these weeks, I got to learn a lot more of the intricacies and the challenges faced during exploratory data analysis. The last week was spent in going through a research paper provided by the mentors and the subsequent realisation of the project by referring to this paper.

References

- A Gentle Introduction to Programming using Python
- What is Machine Learning?
- Python Tutorial
- Numpy
- Pandas
- What is Exploratory Data Analysis? Part 1
- What is Exploratory Data Analysis? Part 2
- Financial Data with Python: yfinance
- Stock Market Prediction using the ARIMA Model Paper
- Stock Market Prediction using ARIMA Model

****