Introduction to Data Analytics

Project Report

Topic 6: Auto-regression analysis with time-series data for future event prediction

Data of submission: 30-11-2021

Group 6:

Bhavesh C: S2019001010034

Navdeep K: S2019001010099

Bhargav K: S20190010083

Khadyothan D: S20190010040

Karthik K: S20190010100

Problem Statement

Reference: STOCK data for the year 2016-2017

a) Let's consider the case of p-th order auto-regression analysis. Obtain the Covariance matrix suitable for p-th order auto-regression correlation analysis.

b) For the given data from stock exchange predict the stock value for the month Dec 2017.

c) Report your prediction with different values of p.

Hint: To solve the above problem, you are free to choose any method reported in the literature.

Understanding the theory

- Auto regression predicts future values based on the past values. So basically the trend of the stocks is being predicted.
- Our data set initially had multiple company stocks , so we extracted 495 rows of the data belonging to one particular stock and applied automatic regression on the stock.
- Based on the data covariance matrix and auto correlation coefficients can be achieved.
- By following Yule-walker equations, we can achieve autoregression coefficients to predict future values based on the past values.
- Since it is a trend analysis rather than using R2 score, we must apply MAPE or SMAPE.
- To counter a situation where we have relatively very close values of MAPE, direction score was considered, ie Direction_score = (number of times the trend of data was predicted correctly by the model) . The idea behind taking this metric is to find how accurate the model is able to predict the trend within the data.

Implementation of the project

- STEP - 1 COLLECTING DATA
  - The given dataset consists of stock prices of multiple stocks at various intervals that include price_open, price_close ets. We have considered the open_price as a benchmark for the model.

- STEP - 2 DATA PRE-PROCESSING
  - A single stock was chosen to build the AR model, i.e 20 microns.
  - Extracted the timestamp and open_price of the stock prices from the entire dataset.
  - Removed NA values and ordered the data according to timestamp.

- STEP - 3 BUILDING AR MODEL
  - Finding covariance matrix.
  - Finding auto-correlation coefficients based on the covariance matrix.
  - Now based on the correlation coefficients, we take the yule-walker equation to find auto-regression coefficients (beta i).
  - $$\begin{pmatrix} 1 & r_1 & r_2 & r_3 & r_4 & \cdots & r_{N-1} \\ r_1 & 1 & r_1 & r_2 & r_3 & \cdots & r_{N-2} \\ r_2 & r_1 & 1 & r_1 & r_2 & \cdots & r_{N-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot \\ r_{N-1} & r_{N-2} & r_{N-3} & r_{N-4} & r_{N-5} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \\ \\ \\ r_N \end{pmatrix}$$

- STEP - 4 FORECASTING
  - Based on the auto regression coefficients, we forecast the p th-order time series

$$x_t = c + \sum_{i=1}^{p} a_i x_{t-i} + \in_t$$

- Gaussian white noise is taken as a random normal gaussian distribution value.
- Plotting the model using ggplot.

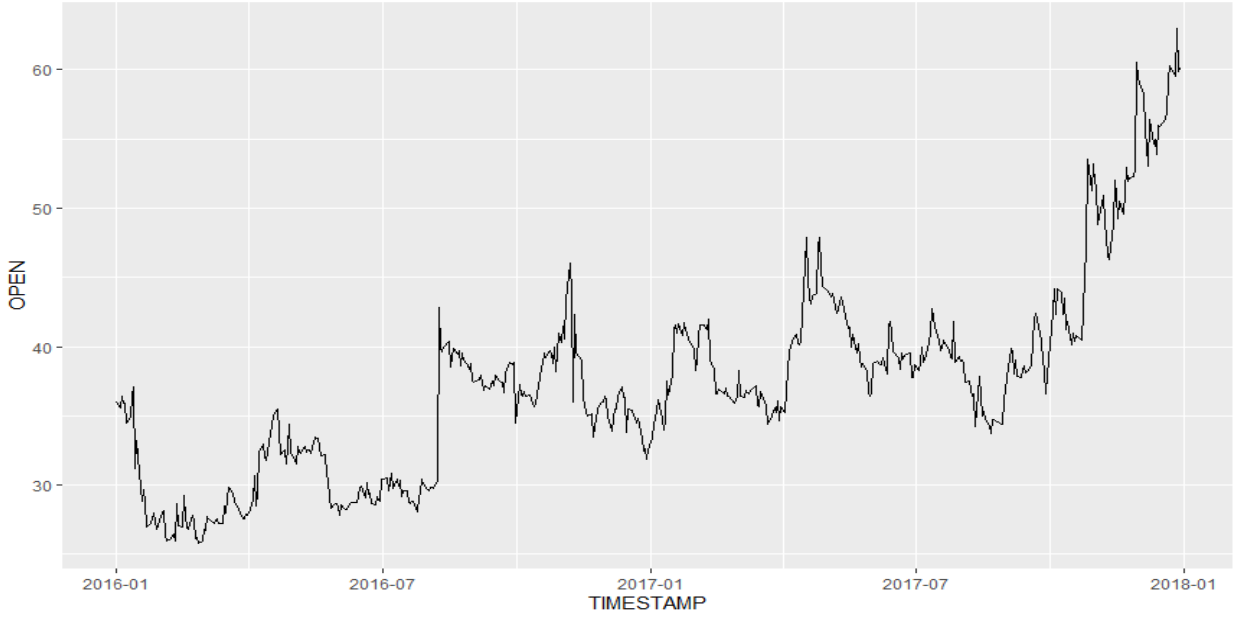
- STEP - 5 ACCURACY METRICS
  - We considered MAPE (Mean Absolute Percentage Error) to evaluate error score of our model

$$MAPE = \frac{\sum \frac{|A-F|}{A} \times 100}{N}$$

  - As the resulting model had close errors, we considered the direction_score metric to predict the better p-value
  - Direction_score = (number of times the trend of data was predicted correctly by the model)

Data and Values

Plot of OPEN vs TIMESTAMP



VALUES

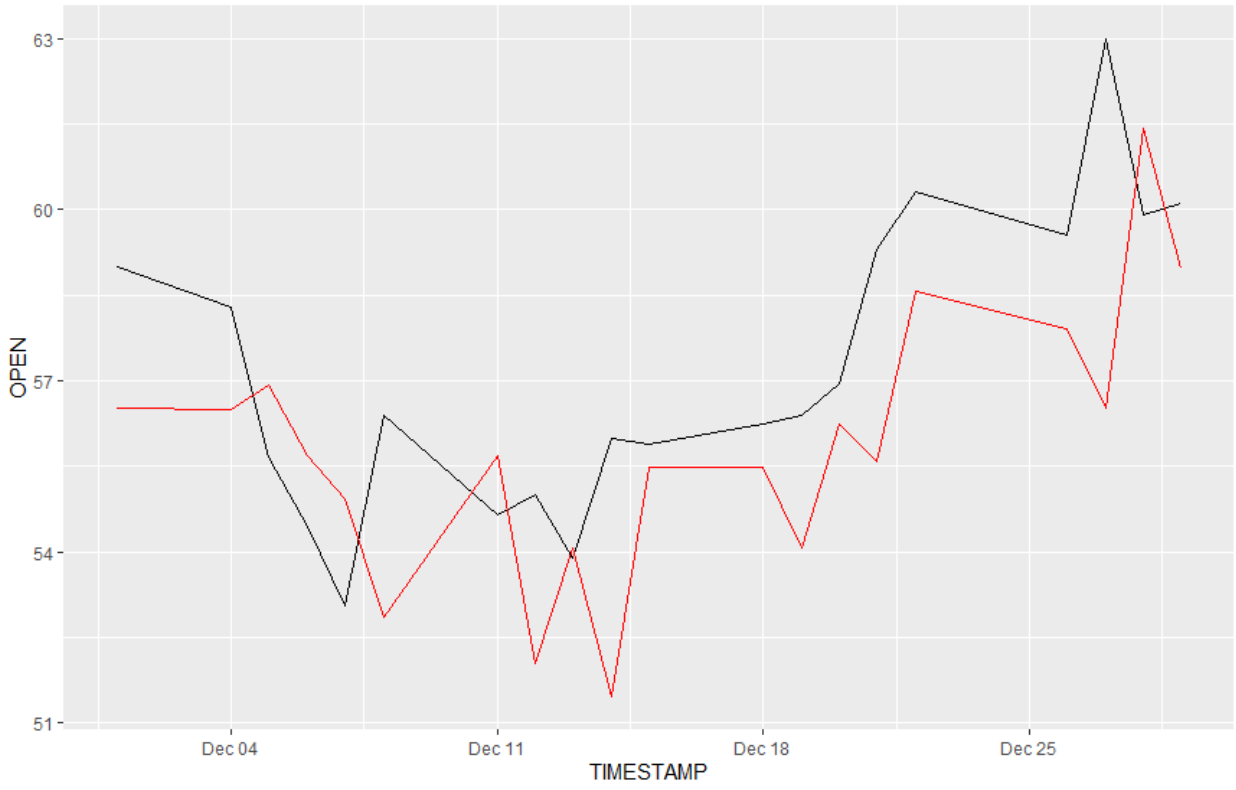
Values	
beta	0.963038392521794
cov	num [1:659] 26.7 26.7 26.1 25.5 25 ...
days	20
dec	num [1:100] 37.5 37 33.8 36 35.3 ...
dec2	num [1:495] 36 35.5 36.5 36 35.9 ...
direction_score	num [1:5] 22.7 20.4 23.1 21.3 22.8
forecast	num [1:20] 58.6 58.7 56.2 54.2 50.8 ...
i	20L
j	10L
mape_arr	num [1:5] 4.34 4.36 4.4 4.3 4.38
mean	37.3779797979798
n	445
p_arr	num [1:5] 9 10 11 12 13
p_test	1
R	0.963038392521794
std	7.21462540934468
sum	43.8178724302979
sum_dir	228
syms	"20MICRONS"
var	52.0508197971619
x	50L
y	num [1:495] 36 35.5 36.5 36 35.9 ...
y_actual	num [1:50] 40.8 40.5 42.6 45.1 47 ...

Experimental results

Finding P-value

- To find the optimal p-value, we ran the AR model over different P-values.
- Finding the optimal pth order, MAPE(mean absolute percentage error) was considered.
- The results showed to have minimum error in the range $P \in (9,13)$
- As the MAPE function was not sufficient to draw conclusion, direction score was considered to find that p = 10, 11 is the optimal value for this AR model.
- NOTE: P-values with $P \in (9,13)$ still have very similar accuracy

Predicting values for the month of december



Conclusion

The AR(1) model has high accuracy in predicting the trends in a model, but fails to meet the seasonal spikes in the data. This makes an AR(i) model with a less i value more vulnerable.

Accuracy metrics are applied to find the optimal p-value, here we got the optimal value as 11.

Since MAPE is a measure of error, high numbers are bad and **low numbers are good**, and it has more value in terms of auto regression since percentage errors help evaluate a proper AR model.

MAPE is 4.362974 for p-value 11.

A MAPE value of 1% - 5% is considered to be a very accurate forecasting.