Stock Forecasting

Rupali Shah | Springboard | Sept 20, 2019

Overview

There are three factors that drive the stock market:

- Fundamental factors drive stock prices based on a company's earnings and profitability from producing and selling goods and services.
- Technical factors relate to a stock's price history in the market pertaining to chart patterns, momentum, and behavioral factors of traders and investors.
- Market sentiment refers to the psychology of market participants.

Technical factors and market sentiment often overwhelm the short run. This project uses Technical factors to predict the stock prices of Oil and Gas companies such as Chevron Corp (CVX) and Exxon Mobil (XON) based on the Stock Price History and the Inflation Rate.

Time Series forecasting is used to predict the stock prices. Depending on the frequency, a time series can be of yearly (ex: annual budget), quarterly (ex: expenses), monthly (ex: air traffic), weekly (ex: sales qty), daily (ex: weather), hourly (ex: stocks price), minutes (ex: inbound calls in a call canter) and even seconds wise (ex: web traffic).

ARIMA, a popular statistical method is used to build the model.

Data Source

The daily stock data for five years (2013-2017) for Chevron Corporation (CVX) is downloaded from the Quandl API.

The 5 year breakdown inflation rate is downloaded from the FRED website in CSV format.

Exploratory Data Analysis

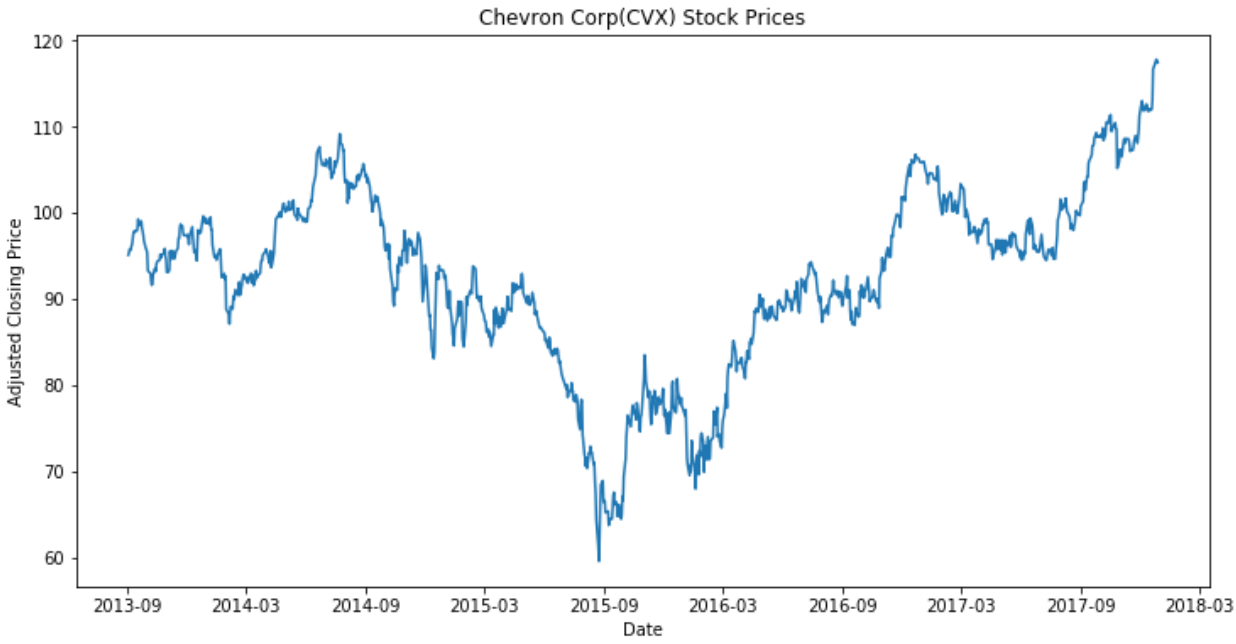
The dataset has 13 columns with Date as the Index column.

- The Date column is the day on which the stock was traded.
- The columns Open and Close represent the starting and final price at which the stock traded on a day.
- High and Low represent the maximum and minimum price of the share for that day.
- Volume is the total number of shares traded in the day.
- The Split column shows if the stock has been split indicating that the number of shares has increased, and the value of each share has decreased.
- The Adjusted Closing price reflects the stock's value after accounting for any corporate actions. e.g. the closing price adjusted after posting a dividend. Thus, it accounts for the newly reduced value caused by the dividend.

Profit/Loss is determined by the Closing Price. Since Closing Price is the raw price, the Adjusted Closing price is considered as the Target variable.

Stationarity Check

- Visual Check
  The plot below indicates that the Time Series is not seasonal and not stationary.



- Statistical Test

  Augmented-Dickey Fuller(ADF) test

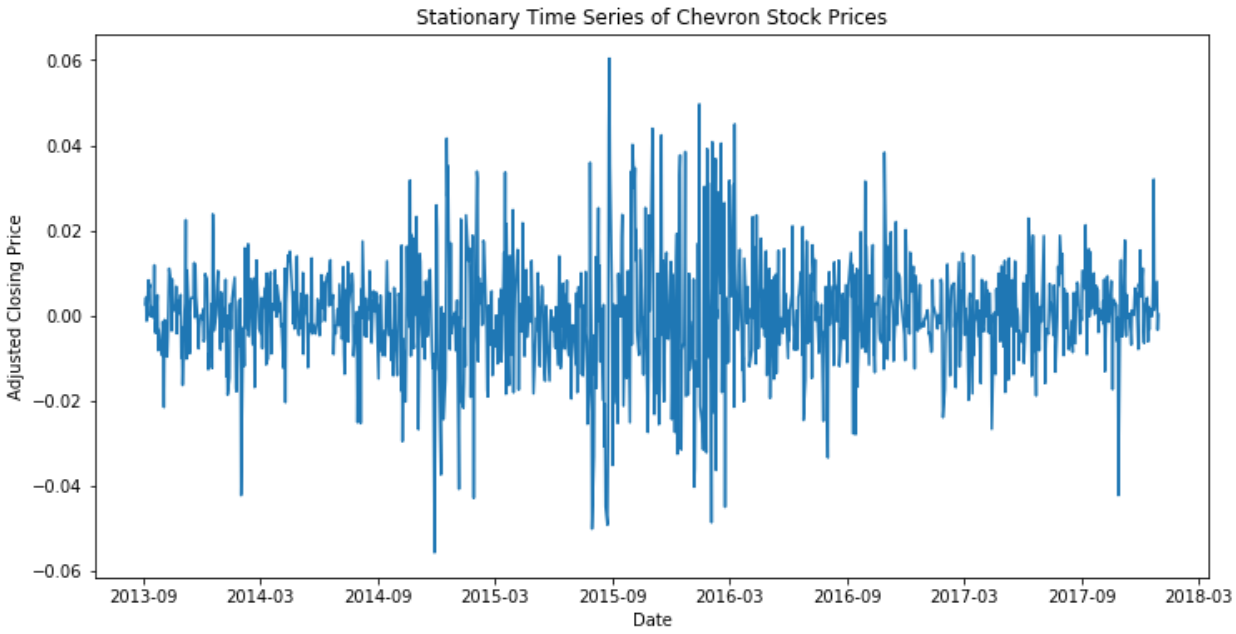
  The null hypothesis of the ADF test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) then the null hypothesis is rejected, and it is inferred that the time series is indeed stationary.

  The ADF test results suggest that Time Series is not stationary - Critical Value is greater than the Test Statistic, and p-value > 0.5.

Test Statistic	-0.948494
p-value	0.771584
Number of lags	0
Number of observations	1089
Critical Value (1%)	-3.436369
Critical Value (5%)	-2.864198
Critical Value (10%)	-2.568185

Data Munging

The Time Series is made stationary by differencing it. This is done by transforming the data to a logarithmic scale to distribute the data normally, and then subtracting the previous value from the current value.



The ADF-test results suggest that the Time Series is stationary since p-vale is less than the significance level, and critical value is less than the Test Statistic.

Test Statistic	-1.78E+01
p-value	3.31E-30
Number of lags	3.00E+00
Number of observations	1.09E+03
Critical Value (1%)	-3.44E+00
Critical Value (5%)	-2.86E+00
Critical Value (10%)	-2.57E+00

Model Building

The model is built using ARIMA, a popular statistical method, which needs a stationary series, and takes three parameters p, d, q.

d – The number of times the series is differenced in order to make it stationary.

p – This is the Auto Regressive term, which refers to the past values for forecasting the next value. This value is determined by the PACF plot.

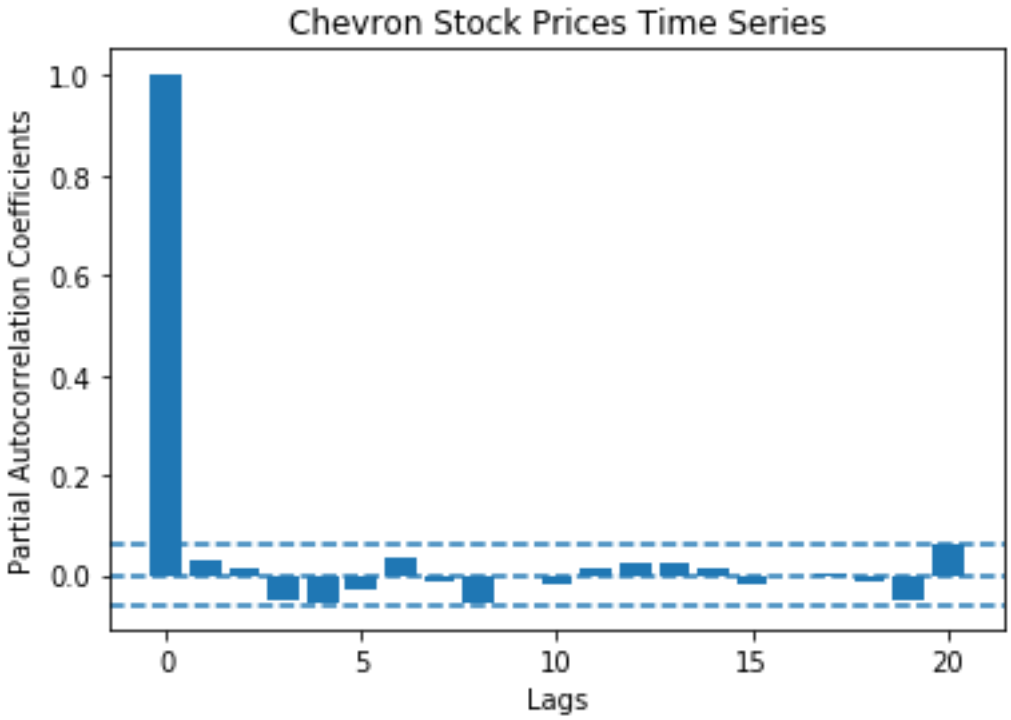
q – This is the Moving Average term, which defines the number of past forecast errors, and is determined using the ACF plot.

Value of d

The Time Series has been differenced once, so the value of d =1

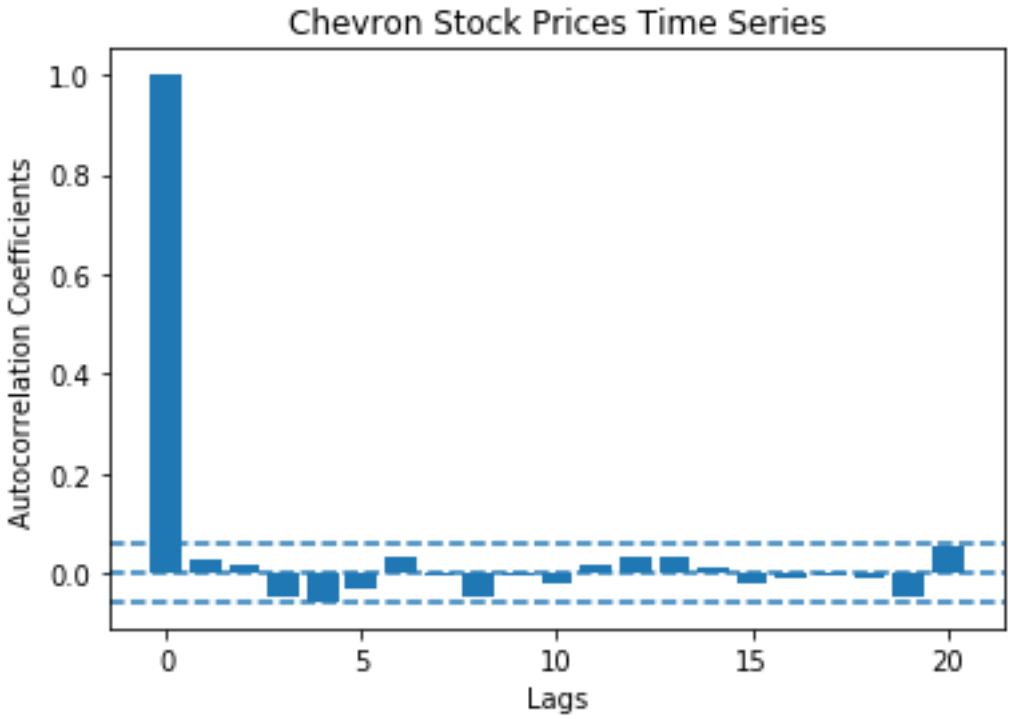
Value of p

The PACF plot is created to determine the value of p.



Value of q

The ACF plot is created to determine the value of q



Fitting the model and Forecasting

The model is fitted with Auto Arima from the Pyramid library.

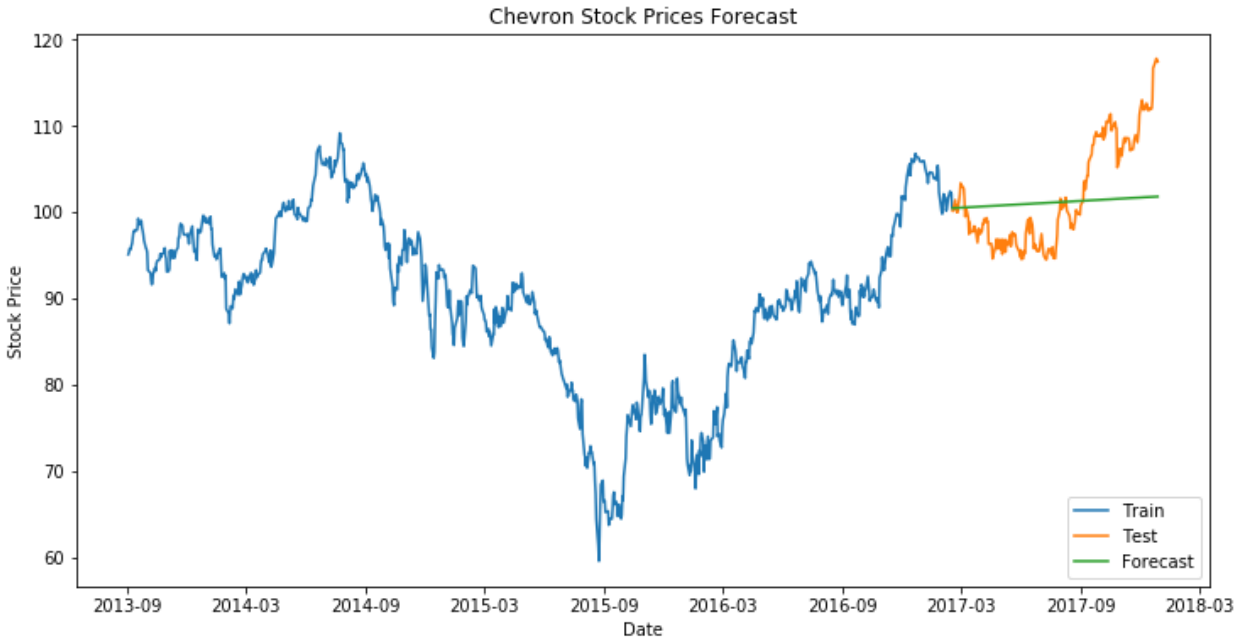
Hyperparameter tuning for the ARIMA model can be quite time consuming. Auto ARIMA automatically selects the best combination of (p,q,d) that provides the least error. Auto ARIMA considers the AIC and BIC values generated to determine the best combination of parameters. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values are estimators to compare models. The lower these values, the better is the model.

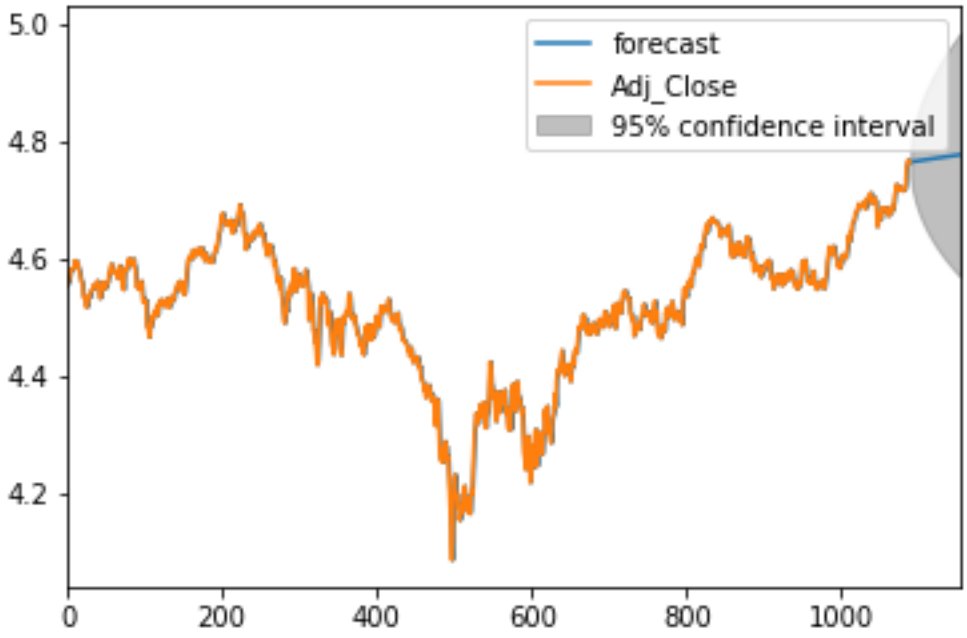
After splitting the data into train and test with an 80/20 split, the model is fitted using auto ARIMA from the Pyramid library. This model is built with the combination (0,1,0) based on the AIC and BIC values.

Forecasting

The predictions are made on the test dataset.

The model gave a root mean squared error of 5.91.




forecast
Adj_Close
95% confidence interval
5.0
4.8
4.6
4.4
4.2
0
200
400
600
800
1000