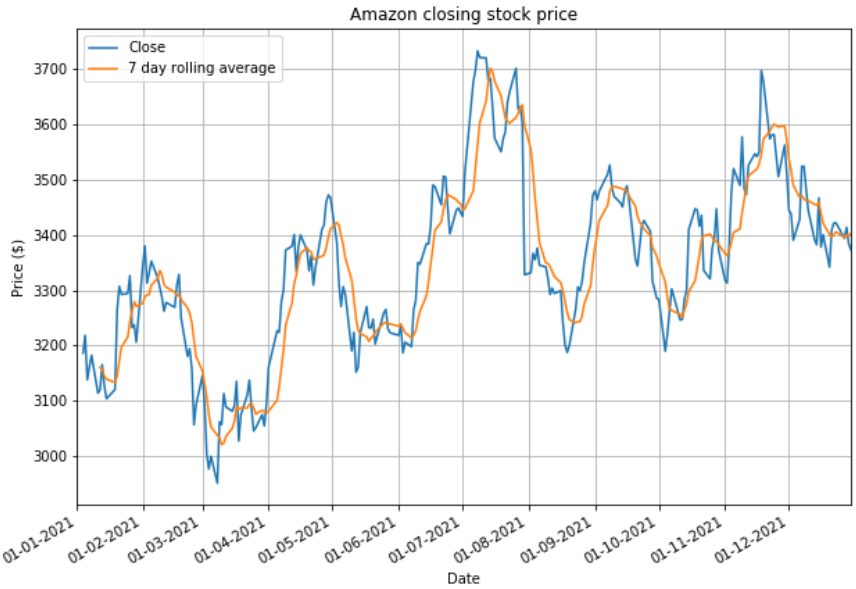
ADFT CA2

Luke Kirwan

February 2022

1

The plot below shows the closing price and 7-day rolling average of Amazon's stock price in 2021.



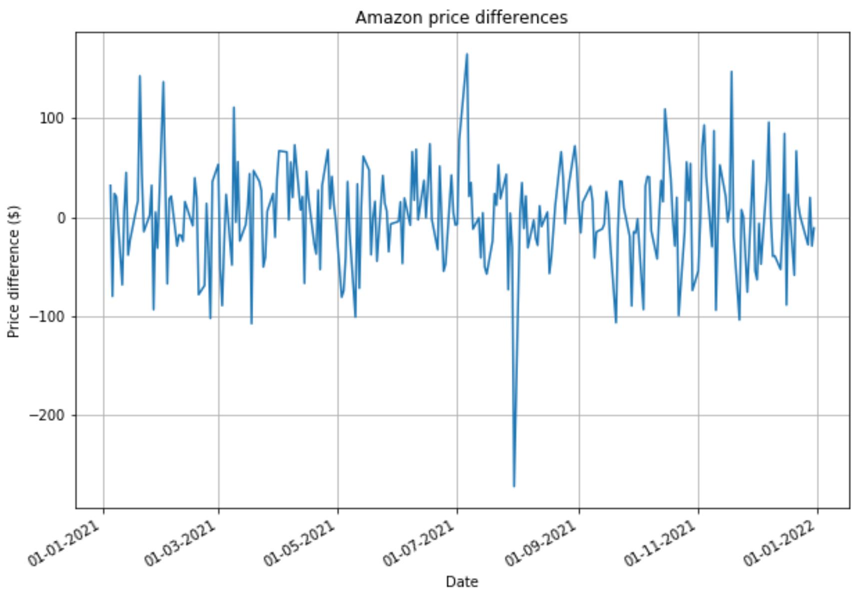
Amazon closes the year at approximately \$3400, roughly \$200 higher than the price it had at the start of the year. This is as a result of the gradual upward trend in the data, which shows that the mean price slowly increased throughout the year. This trend implies that the series is not stationary. There also appears to be a cyclical pattern of peaks. This pattern appears to be cyclical, and not seasonal, because the length of time between peaks is fairly unpredictable.

2

Stationary time series have statistical properties which don't change no matter which sub-section of the series you consider. So, properties such as mean and variance remain constant throughout the series.

Amazon's stock price doesn't appear to be stationary. Firstly, there is a trend of increasing prices, suggesting the mean price is increasing throughout the series. Secondly, there appears to be some sort of cyclical pattern, shown by regular but inconsistently spaced peaks. These observations make it unlikely that the series is stationary.

We can use differencing to transform this series into a stationary time series [1]. From the graph below, you can see that the trend has been removed. The transformed series is much more likely to be stationary than the original series of raw price values.



3

Autocorrelation is the correlation of a time series with itself at a different point in time, as a function of the lag between the observed times. For a time series which is closely correlated with recent values, but not related to distant values, the autocorrelation will shrink as the time lag increases. We can use the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to measure the autocorrelation of a series. [1]

The ACF is used to determine the Pearson correlation coefficient between values at different times. For a given lag k, and time series $y_1, ..., y_n$, the correlation coefficient r_k is:

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

[2]

One issue with the ACF is that it is dependent on the intermediate values in the series. The PACF focuses on the direct correlation between two values of a time series, removing the affect of these intermediate values. Calculating the PACF is more complicated than the ACF, however it can give us a better insight into the dependence of a time series on its past values, since the PACF will disappear much faster than the ACF for time series which are heavily dependent on recent values. The PACF can be used to determine the most appropriate order for autoregressive models.

For a given lag k, the partial autocorrelation of y_t with y_{t-k} is:

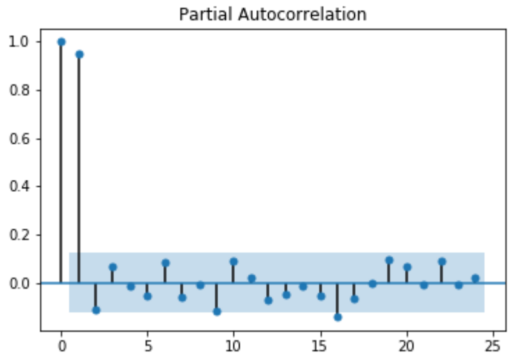
$$\phi_k = \frac{\text{Cov}(y_t, y_{t-k} | y_{t-k+1}, ..., y_{t-1})}{\sqrt{\text{Var}(y_t | y_{t-k+1}, ..., y_{t-1}) \cdot \text{Var}(y_{t-k} | y_{t-k+1}, ..., y_{t-1})}}$$

[3]

In this formula, the conditional dependence on the intermediate values $y_{t-k+1,...,y_t}$ is clear.

4

First, we plot the PACF of the closing stock price, to identify autocorrelation and help choose a lag for the autoregressive (AR) model.

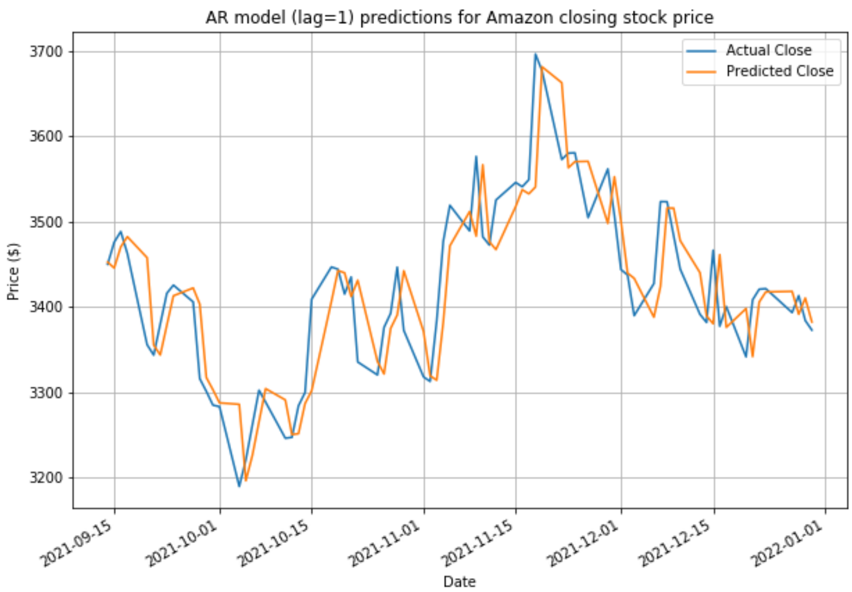


It is clear that the PACF drops off very quickly after a lag of 1, and the strongest direct correlation is between a value and the value preceding it. So, we choose a lag of 1 when fitting our AR model, which gives a regression model of the form:

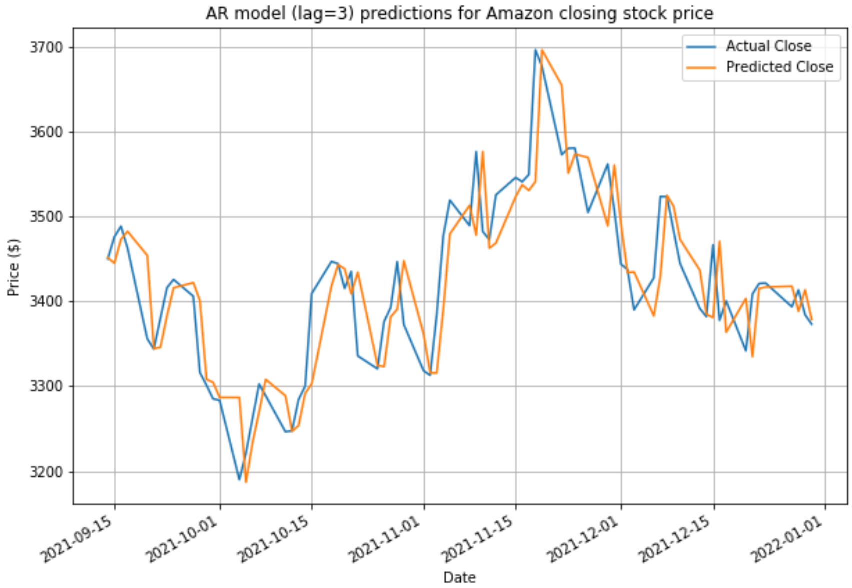
$$Y_t = \theta Y_{t-1} + c$$

Where Y_t is the closing price at time t. For larger lags this formula would include more previous time values.

We use 70% of the available data to train the model, during which we find the optimal values for the parameters θ and c using linear regression and least squares. We then use the formula above to make predictions on future prices, which are plotted alongside the actual prices below.



The AR model is a fairly good fit. However, it is slightly delayed in its predictions, which can be seen by the rightwards shift of the predicted curve. The root mean squared error (RMSE) for this model is approximately 54.3. Interestingly, the RMSE for a model with a lag of 2 is approximately 51.8, and for a lag of 3 is 51.6. The graph below shows that a choice of 3 for the lag results in a marginally better fit on the test data. When we increase lag above 3, the RMSE creeps back up. This suggests a lag of 3 is actually the best choice, despite the inferences made from the PACF plot.



5

One of the key strengths of AR models is their simplicity. They are easy to use and quick to learn, and there are tried and tested techniques for computing the optimal parameters.

Another benefit is that we can create a model for predicting future values based purely on past values. This is a powerful tool and can be used as a 'base' for more complex models, which might incorporate external information. However, this assumption of past values being useful in predicting future values can be problematic and lead to wildly inaccurate predictions in certain scenarios.

Another weakness of AR models is that they only take the current error into account, and not past errors. Thus, if there is a correlation between the error terms, an AR model will not use this information. One way we can overcome this weakness is by using Moving Average (MA) models. MA models use past errors in a regression like model, and can be combined with AR models to make more powerful models. [1]

A final problem with AR models is that it can be cumbersome to retrain them as new data becomes available. However, this problem is not unique to AR models.

References

[1] Nielson, Aileen, *Practical Time Series Analysis : Prediction with Statistics and Machine Learning* 2019, First ed. Sebastopol, CA: O'Reilly Media.

[2] www.itl.nist.gov *1.3.5.12. Autocorrelation* `https://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm`

[3] www.real-statistics.com *Partial Autocorrelation Function* `https://www.real-statistics.com/time-series-analysis/stochastic-processes/partial-autocorrelation-function/`