Weijia Wang

This dataset includes columns “open”, “high”, “low”, “close”, “adj close” and “volume” from Nov 16, 2020, to Nov 12, 2021. The first step is to explore the data. The first chart (Fig. 1) shows the date on the x-axis and close prices on the y-axis. The second graph (Fig. 2) plots the distribution of close prices using kernel density estimation (KDE). Based on the plot, there are two peaks, which means the density is the highest when the prices are around $130 and $150.

Then the second step, we test for stationarity. Since stock markets are changing significantly, we want to use a rolling window can assess the stability over time. In this case, the rolling window is 5, which means the value on Nov 20, 2020, after rolling is the mean of the first 5 values in the original data. In Fig. 3, the red line represents the mean after rolling and the black line represents the standard deviation. Then we apply the Augmented Dickey-Fuller Test (ADF test) to test whether the time series data is stationary or not. Since the p-value is 0.511 and greater than the significant value (0.05), we can reject the null hypothesis, which means the time series data is non-stationary.

In the third step, we use the seasonal\_decompose function to decompose the close price into “seasonal”, “trend” and “resid”. The trend component is a long-term increase or decrease which might not be linear. Sometimes the trend might change direction as time changes. The seasonal component is a series that exhibits regular fluctuations based on the season. Seasonality is always of a fixed and known period. The residual component is what is left over after fitting a model.

In Fig. 5, we split the time series data into testing and training sets. Using auto\_arima function to fit the ARIMA model. There are three different terms, p, d, and q. Term “p” is the order of the auto-regressive (AR model). Term “d” is the degree of differencing. Term “q” is the order of the moving average (MA) model. There are four output charts. The *standardized residual* shows the residual errors appear to have a uniform variance and fluctuate around a mean of zero. The *histogram plus estimated density* shows the density plot on the top right suggests a normal distribution with a mean of zero. The *Normal Q-Q* means the red line should be perfectly aligned with all the dots. Any significant deviations would indicate a skewed distribution. The *Correlogram*, also known as the ACF plot, represents the residual errors are not autocorrelated. Any autocorrelation would imply that residual errors have a pattern that isn’t explained by the model. The ARIMA model assigned the value p=0, d=1, and q=0. Using the p, d, and q values we got from the previous step, we plot the ARIMA model with training data and order = (0, 1, 0).

The log-likelihood value represents the maximum likelihood estimation. The higher the log-likelihood, the better. AIC stands for Akaike’s Information Criterion. It helps evaluate the strength of the model. AIC deals with both the risk of overfitting and the risk of underfitting. The lower the AIC value is, the better the model is performing. BIC stands for Bayesian Information Criterion. It is very similar to AIC and the lower BIC value is preferred. The HQIC stands for Hannan-Quinn information criterion, and it can be used for feature selection.

For the ‘coef’ column, ar.L1 refers to the autoregressive term with the lag of 1, ar.LA means the lag of 2. Ma.L1 and ma.L2 indicate the ‘moving average’ term with a lag of 1 and 2. The “std err” shows the estimate of the error of the predicted value. It indicates how strong is the effect of the residual error on the estimated parameters. The ‘z’ is the standardized coefficient which equals to the values of ‘coef’ divided by ‘std err’. The ‘P>|z|’ column is the p-value of the coefficient. Usually, p-values should be lower than the given threshold (0.05). The last two columns represent the confidence intervals between 2.5% and 97.5%.

Then, we build and forecast the ARIMA model (Fig. 6) and plot stock prices on the test dataset with a 95% confidence level. Based on the graph, the predicted values and actual values lie within the 95% confidence band. However, the predicted values are below the actual values, which means we can add a constant to increase the accuracy. After trying the new values (p=2, d=2, and q=0), all the p-values are less than 0.05.

In the last step, we calculate the MSE, MAE, RMSE, and MAPE. MSE stands for Mean Squared Error, which measures the average of squared forecast error values. MAE stands for Mean Absolute error which represents the absolute difference between the actual and forecast values in the dataset. It measures the average of the residuals in the dataset. RMSE stands for root mean squared error, which means the square root of mean squared error. It measures the standard deviation of residuals. Mean absolute percentage error (MAPE) measures the prediction accuracy of a forecasting method in statistics. In this case, the MAPE value is 0.00482, which means around 0.48%, the model is 97.5% accurate in predicting the next 15 observations.