Penny Stock Prediction

By: Richard Broyles

Introduction

The stock market has been an integral part of the U.S. economy for well over 200 years. The fate of our whole economic system of capitalism lives, and sometimes dies, with the stock market. Since the beginning of the technological age, investors have been wanting to know where they should put their money in the stock market to make themselves financially secure. Since the advent of the internet, stock market information has been readily available to not only financial advisors and stock market professionals, but to any trader who wished to trade on the stock market.

The goal of this project is to predict the price of a couple of penny stocks from various sectors of the stock market. A penny stock is defined by the SEC (Securities and Exchange Commission) as a stock that has a value of $5 per share or less. For this project, we will be looking at a couple of penny stocks, Inuvo Inc. and Biolase Inc.

Inuvo Incorporated was founded in Little Rock, Arkansas and is a technology company that develops and sells information technology solutions that identify and message online audiences across various platforms. Their platform also allows advertisers and publishers to buy and sell advertising space in real time. The company's marketing channels consist of websites, social media, blogs, public relations, trade shows, and conferences (http://finance.yahoo.com/quote/NUV?p=INUV).

Biolase, Inc., develops, manufactures, markets and sells laser systems for dental practitioners and their patients in the U.S. and internationally. Their laser system allows dentists to perform a range of minimally invasive dental procedures. Biolase was founded in 1994 and is currently headquartered in Irvine, California. (http://finance.yahoo.com/quote/BIOL?p=BIOL)

The data for this project is captured in real time using the DataReader library that allows real-time data to be collected from Yahoo! Finance, Google, and various other financial sectors. It is a part of the Pandas library. There is 10 years of data that is collected. There are only six variables for each stock:

1) High – The highest value of the stock in the day's trading.
2) Low – The lowest value of the stock in the day's trading.
3) Open – The value of the stock when the market opened for that day.
4) Close – The value of the stock when the market closed for that day.
5) Volume – The number of shares that were traded on that day.
6) Adjusted Close – This value is used in examining historical returns of a stock. This is calculated after business hours and considers dividends, stock splits, and any movement that occurs after the market closes.

Since we are looking at data over a time period of ten years, we will be using a time series analysis to explore the data and we will be using the ARIMA (Auto Regressive Integrated Moving Average) model to predict future values of these two stocks.

Exploratory Data Analysis

Because the data is live (the data is collected every day from Yahoo! Finance after the close of the market), data cleaning is not necessary. Each stock has over 2,500 data points to analyze, which is equivalent to 10 years of market data.

The first value that was looked at was the history of the stock's adjusted close, which is shown in Figure 1.

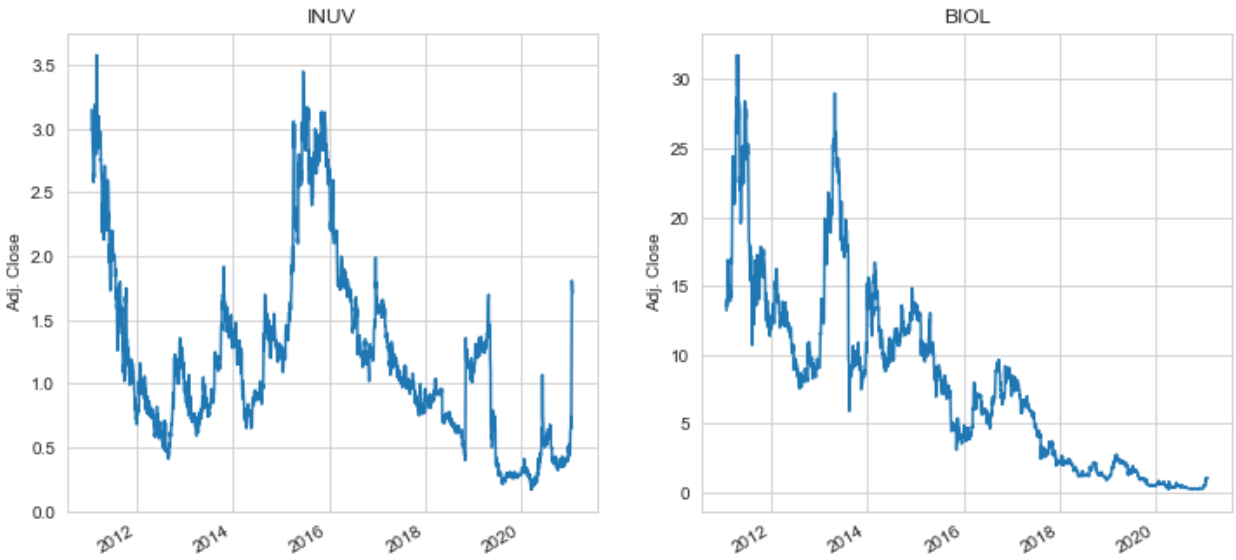

Figure 1 – History of the stock's adjusted close

The first observation about the Inuvo stock is the U-shaped chart at the beginning of the chart, which is shown between the years of 2012 to early 2016, which shows a very steep drop in the value of the stock. The Biolase stock has been dropping in value steadily since late 2013 to 2020 when its value approached zero.

The best way to analyze a stock's performance would be to compare the mean value of the adjusted close of the stock versus the moving average over a certain period. The moving average is used to smooth out the price data over a period by creating a constantly updated average price. There are two types of moving averages, simple and exponential. For this analysis, the simple moving average was used and is defined as:

$$SMA = \frac{A1+A2+\ldots+An}{n}$$

where A is the mean in period n and n is the number of time periods (http://www.investopedia.com/terms/m/movingaverage.asp). For this analysis, the moving average of both stocks were computed over a 10-day, 20-day, and 50-day period. This is shown in Figure 2.

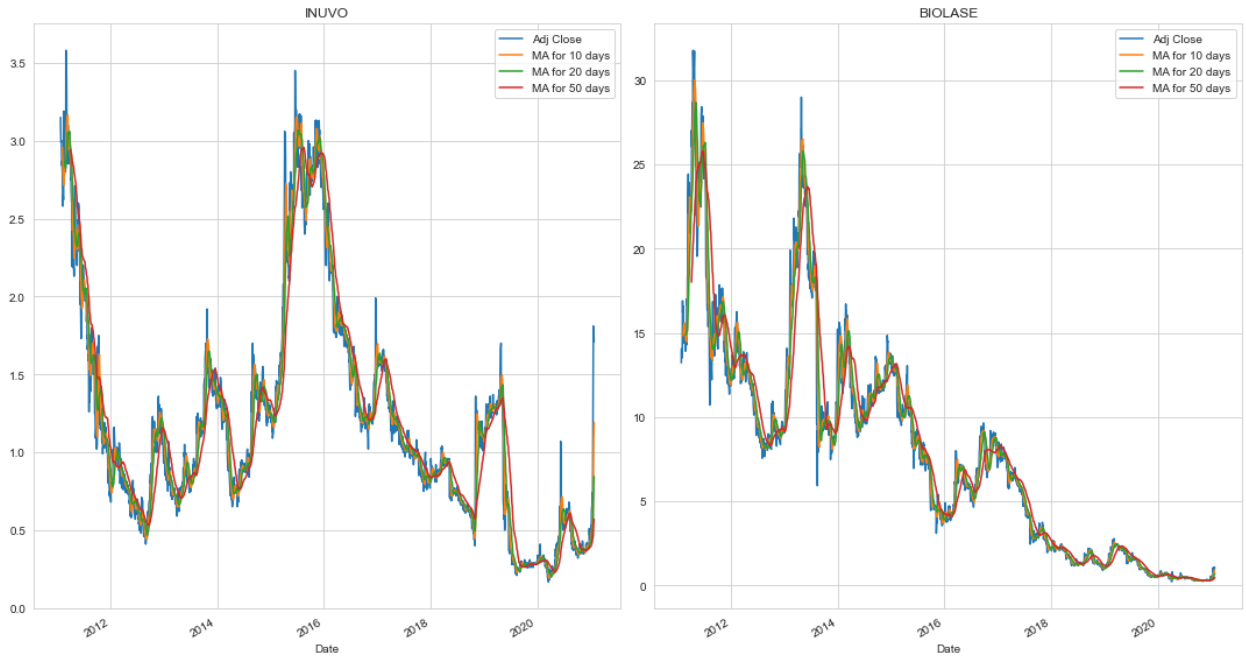

Figure 2 – Moving averages for both stocks

Since the data is live, this figure is as of the close of the market on January 25, 2021.

Another technical measure is the stock's daily return. The daily return measures the dollar change in a stock price as a percentage of the previous day's closing price. This is calculated by subtracting the stock's close the day before from the close on the previous day, then taking the value and dividing it by the closing value on the previous day and multiplying by 100. For example, a stock you own closed yesterday at $50.00 per share and at $48.20 on the previous day. Subtracting the close on yesterday from the previous day yields a value of $1.80. Dividing $1.80 by $48.20 and multiplying by 100 gives a value of 3.73%, which means the value of the stock gained in value by 3.73%.

The daily return for the stocks used is shown in Figure 3.

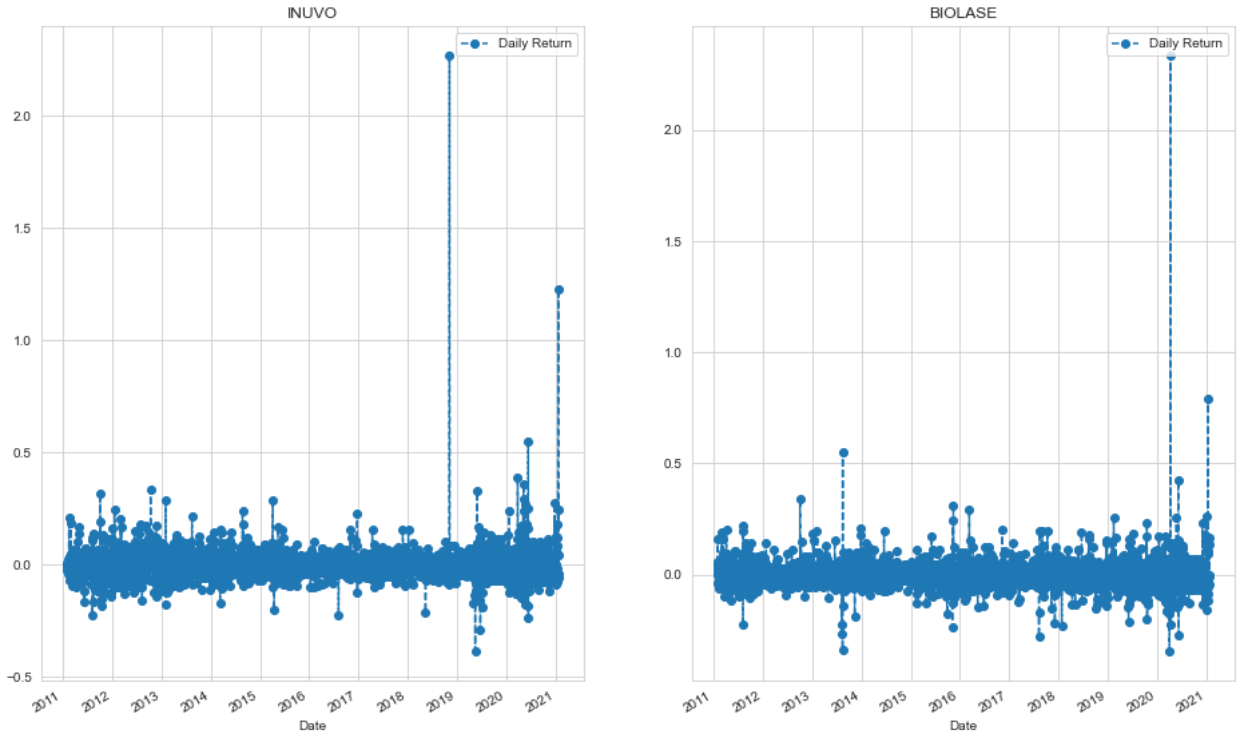

Figure 3 – Daily return for Inuvo and Biolase

A final metric that concerns investors is the metric of risk versus return. All investments carry some risk, which is the value of the stock can go down. Usually, small losses are acceptable as it is often associated with the ebb and flow of the market; large losses are usually the cause for panic in an investor as it represents a large loss in their investment. Investors want to have the highest possible return value for their investment. Since risk and return are highly correlated, higher returns usually involve higher risk.

In this project, we are evaluating two penny stocks, whose risk is their current value. Any substantial moves with these stocks and the risk/return value increases or decreases greatly in terms of their percentage, but the value in terms of real dollars is exceedingly small. These two stocks are a prime example of this. See Figure 4.

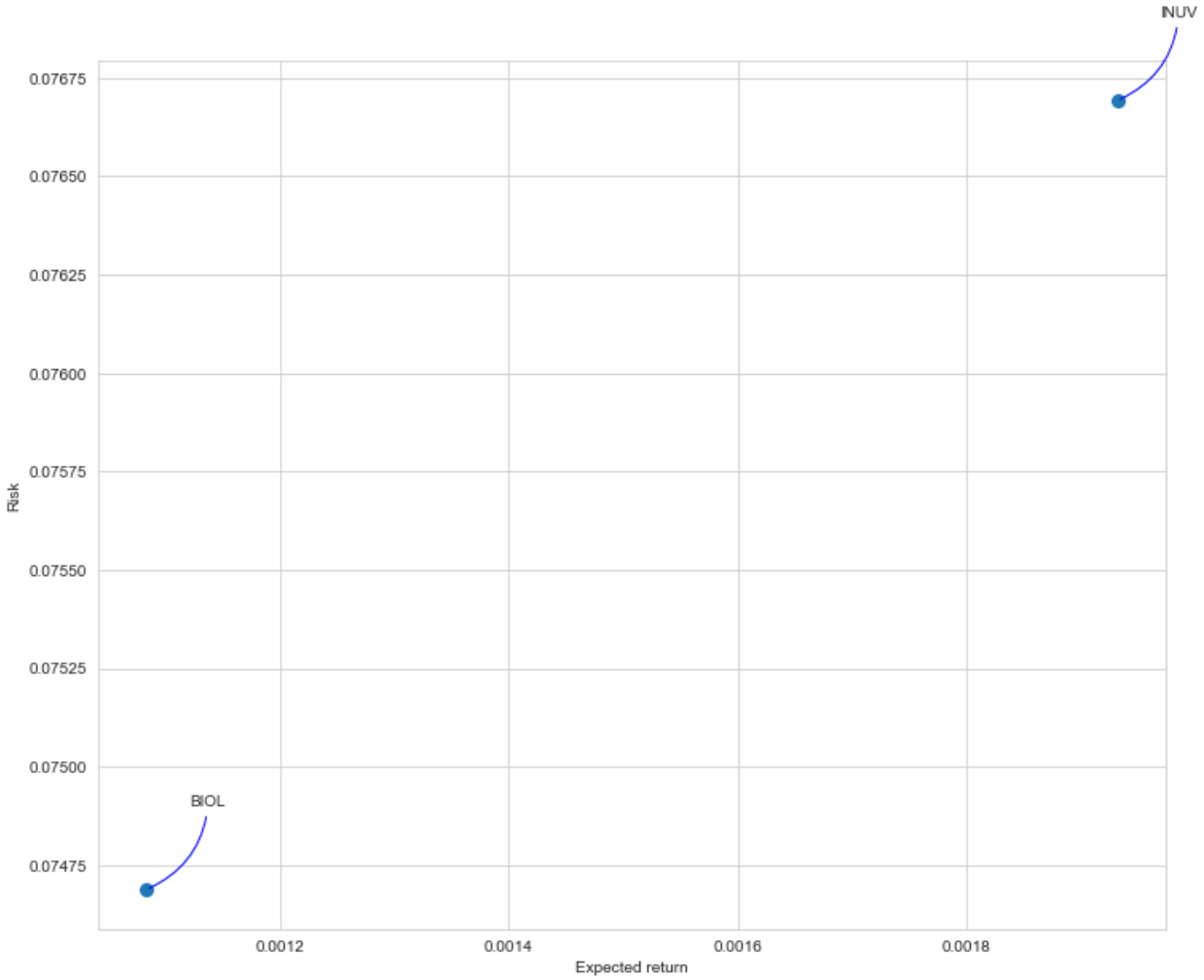

Figure 4 – Risk/return values for each stock. Please note that this data is as of the close of the market on January 25, 2021.

Note that the Inuvo stock currently has the highest risk/return value versus Biolase which has the lowest risk/return value. This is because the Inuvo stock currently has had more moves in the market, mainly to the positive side, while Biolase has been steady. Also, note the exceedingly small value of the scale used in the plot. While these numbers are small, one change in the value of the stock can cause some huge changes in this plot, such as two values exchanging their places in the plot.

Modeling

A time series consists of three systematic components and one non-systematic component. The systematic components are:

1) Level – the average value of the series
2) Trend – increasing or decreasing value of the series
3) Seasonality – repeating short-term cycle of the series.

The non-systematic component is the noise, which is the random variation of the series. The first step in the modeling process is to see if the time series is stationary or not since a time series only works with stationary data. We will be using the ADF (Augmented Dickey-Fuller) test to verify our data is stationary.

The ADF is used to determine whether our time series has the presence of a unit root in the series, thus telling us whether the series is stationary. The null and alternate hypothesis of this test is:

Null hypothesis (H_0) – Series has a unit root ($\alpha = 1$)

Alternate hypothesis (H_1) – The series has no unit root.

If we fail to reject the null hypothesis, the series is non-stationary, which means the series can be linear or difference stationary. If both the mean and standard deviation are flat (constant mean and constant variance), the time series becomes stationary. Figure 5 is the ADT test for the Inuvo stock and Figure 6 is the result of the Dickey-Fuller test.

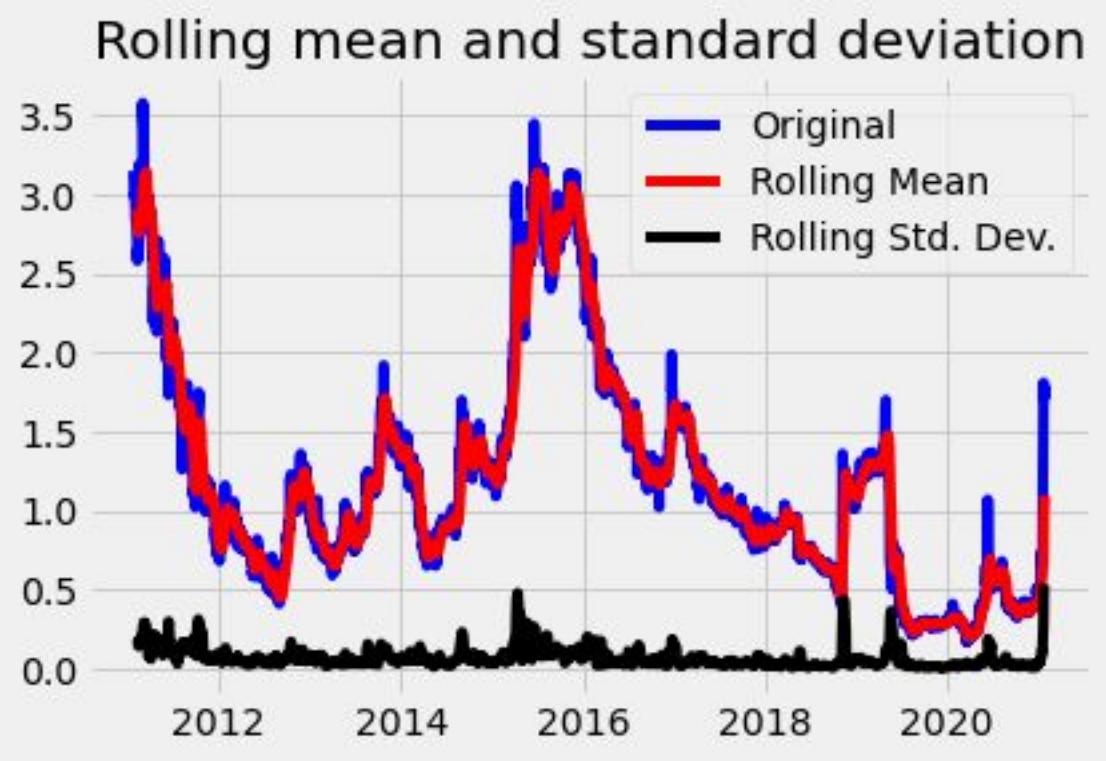

Figure 5 – ADT test plot for Inuvo

Test Statistics	-3.064921
p-value	0.029260

Number of lags used	1
Number of observations	2515
Critical value (1%)	-3.432953
Critical value (5%)	-2.862690
Critical value (10%)	-2.567382

Figure 6 – Inuvo stock Dickey-Fuller test results

Since the p-value is less than 0.05, we reject the null hypothesis, and we conclude that there is no unit root in this series. Figure 7 shows the ADT plot for the Biolase stock and Figure 8 has the results of the ADT test.

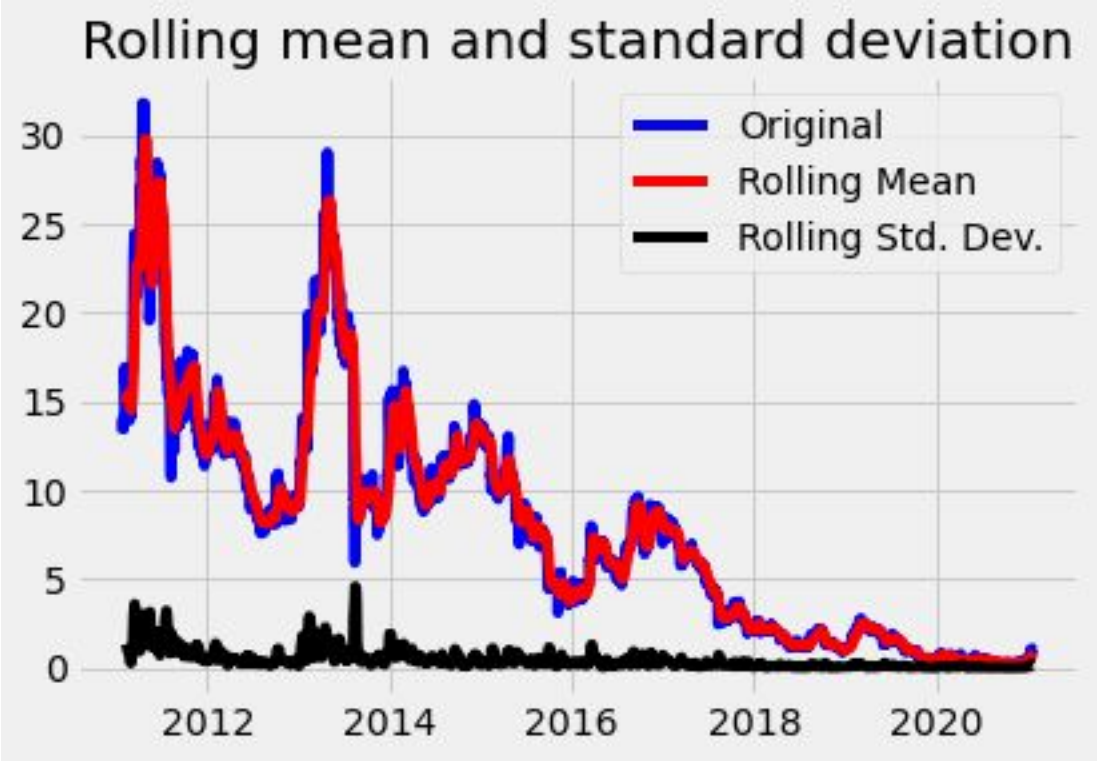

Figure 7 – ADT plot for Biolase stock

Test Statistic	-2.054780
p-value	0.263082
Number of lags used	26
Number of observations	2490
Critical value (1%)	-3.432979
Critical value (5%)	-2.862701
Critical value (10%)	-2.567388

Figure 8 – Biolase stock ADT test results

Since the p-value is greater than 0.05, we cannot reject the null hypothesis and the critical values are less than the test statistic. Given this information, we can conclude that the series is non-stationary.

To further build our model, we will need to separate the trend and seasonality, which allows the resultant series to become stationary. Figure 9 is the Inuvo decomposition and Figure 10 is the Biolase decomposition.

Figure 9 – Inuvo stock seasonal decomposition

Figure 10 – Biolase seasonal decomposition

If we look at the Inuvo stock, we are starting to see a trend upwards, as supported in the current trend in the stock price as well as looking at Figures 2, 3, and 5. The Biolase stock currently has a flat trend meaning that the residuals are near zero.

Before creating the test and training data sets, we need to reduce the magnitude of the closing price and reduce any rising trends in the series. This might not make a lot of sense for penny stocks, but it is necessary to allow the models to make some sort of sense. The log of the closing price data frame was taken and then a 12-month rolling average of the mean and standard deviation was taken for each stock as shown in Figures 11 and 12.

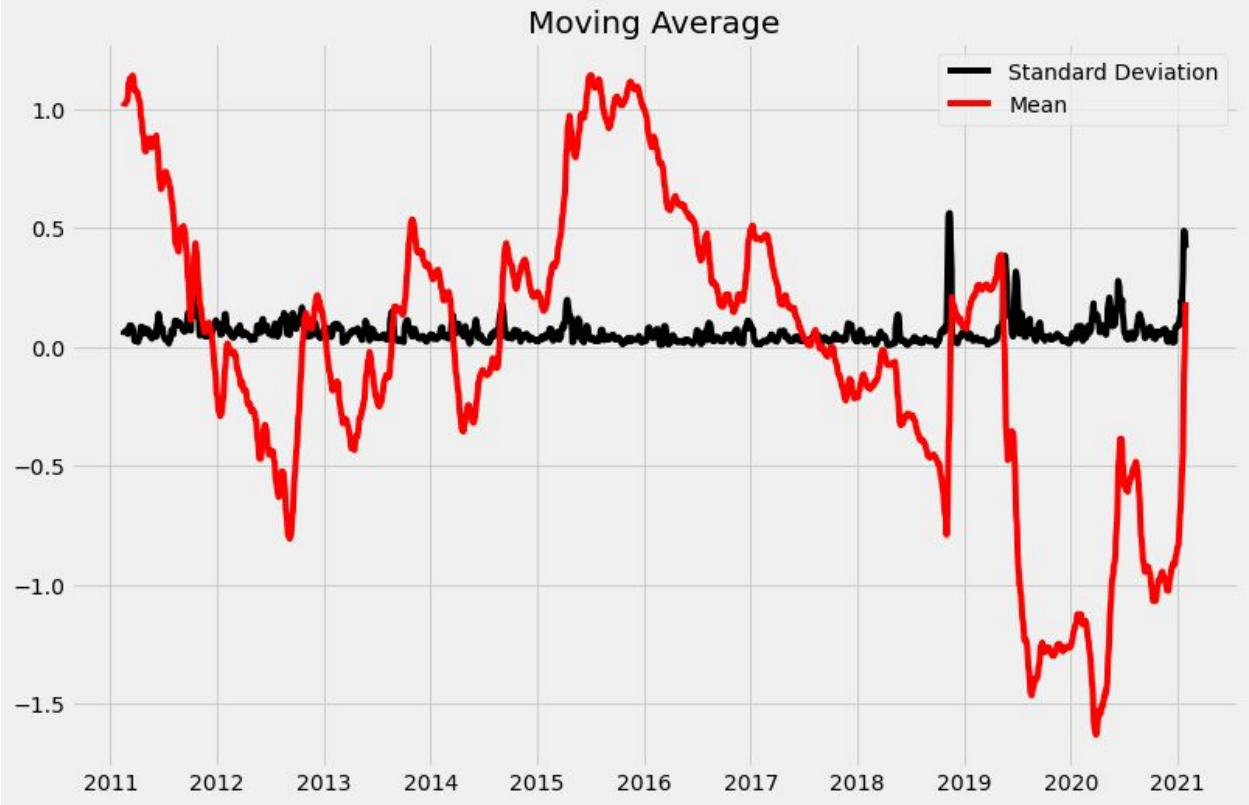

Figure 11 – Log normalization of Inuvo.

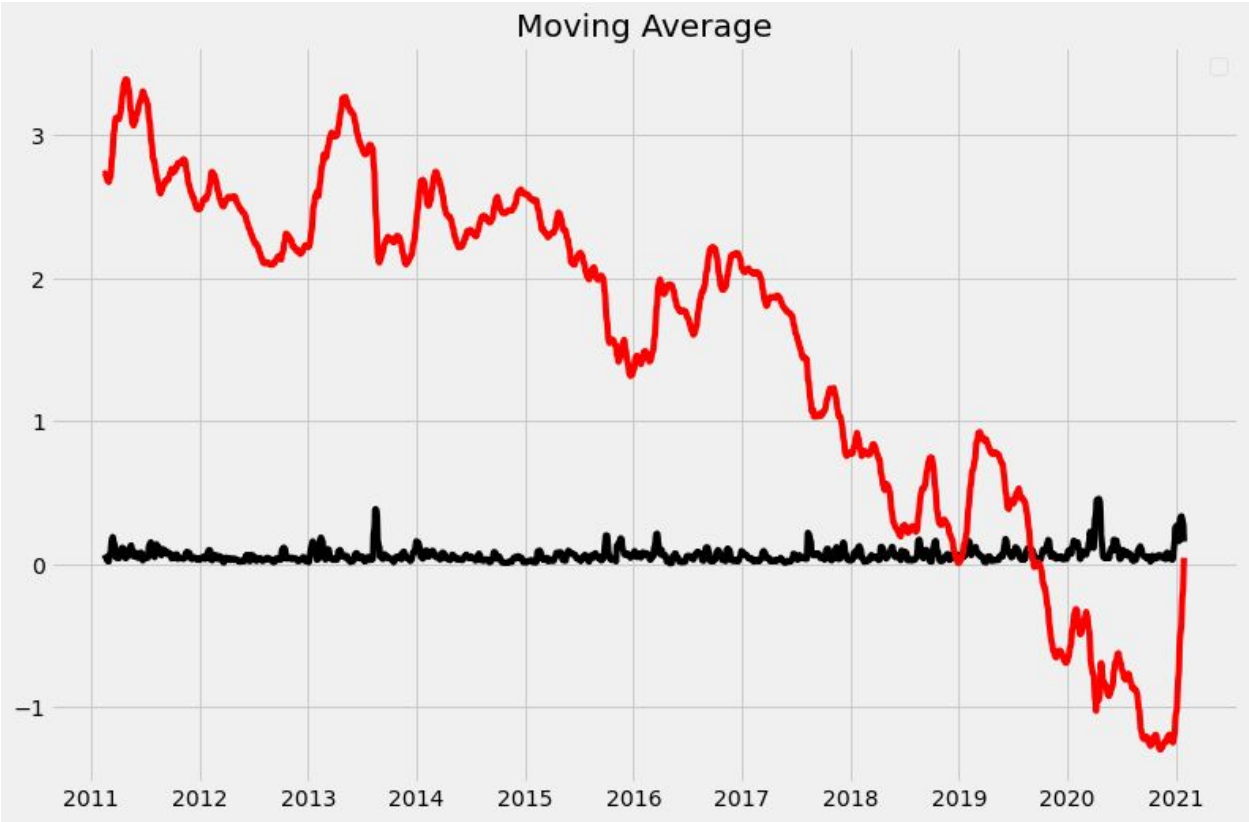

Figure 12 – Log normalization of Biolase.

Prediction

We are using the ARIMA model to predict the future values of these two stocks. The Auto ARIMA function was used to discover the optimal order for the model. This function seeks to identify the most optimal parameters (p, d, and q) for an ARIMA model and returns a fitted ARIMA model. This function works by conducting differencing tests (Kwiatkowski-Phillips-Schmidt-Shin, Augmented Dickey-Fuller, or Phillips-Perron tests) to find the order of the differencing, defined as d, and then fit those models within the ranges defined as start_p, max_p, start_q, and max_q ranges. The values for these stocks are shown in Figures 13 and 14.

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-6484.366, Time=0.23 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-6494.372, Time=0.25 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-6494.899, Time=0.82 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-6485.712, Time=0.08 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-6494.691, Time=0.89 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=-6493.856, Time=0.35 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-6494.049, Time=0.66 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-6496.126, Time=0.10 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-6495.841, Time=0.08 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=-6495.040, Time=0.10 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=-6495.615, Time=0.13 sec
```

```
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=-6495.050, Time=0.31 sec

Best model:  ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 4.028 seconds
                               SARIMAX Results
==========================================================================
====
Dep. Variable:                        y   No. Observations:
2261
Model:               SARIMAX(0, 1, 1)   Log Likelihood
3250.063
Date:                Fri, 29 Jan 2021   AIC
-6496.126
Time:                        03:34:23   BIC
-6484.680
Sample:                             0   HQIC
-6491.949
                                - 2261
Covariance Type:                  opg
==========================================================================
====
                 coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------
----
ma.L1         -0.0747      0.015     -4.919      0.000      -0.105
-0.045
sigma2         0.0033   1.52e-05    217.238      0.000       0.003
0.003
==========================================================================
=========
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):
686881.92
Prob(Q):                              0.99   Prob(JB):
0.00
Heteroskedasticity (H):               1.19   Skew:
3.82
Prob(H) (two-sided):                  0.02   Kurtosis:
88.07
==========================================================================
=========

Warnings:
[1] Covariance matrix calculated using the outer product of gradients
(complex-step).
```

Figure 13 – Auto ARIMA modeling for training data for Inuvo.

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-6909.944, Time=0.20 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-6915.483, Time=0.14 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-6915.873, Time=0.15 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-6910.250, Time=0.07 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-6914.553, Time=0.34 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=-6915.009, Time=0.65 sec
```

```
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=-6913.557, Time=0.42 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-6915.961, Time=0.13 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-6914.569, Time=0.16 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=-6915.004, Time=0.39 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=-6915.586, Time=0.07 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=-6913.575, Time=0.48 sec

Best model:  ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 3.211 seconds
                               SARIMAX Results
==============================================================================
====
Dep. Variable:                          y   No. Observations:
2261
Model:               SARIMAX(0, 1, 1)   Log Likelihood
3459.981
Date:                Fri, 29 Jan 2021   AIC
-6915.961
Time:                        03:34:27   BIC
-6904.515
Sample:                             0   HQIC
-6911.785
                                - 2261
Covariance Type:                  opg
==============================================================================
====
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
ma.L1         -0.0598      0.010     -5.842      0.000      -0.080
-0.040
sigma2         0.0027   3.62e-05     75.614      0.000       0.003
0.003
==============================================================================
=========
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):
6429.79
Prob(Q):                              0.98   Prob(JB):
0.00
Heteroskedasticity (H):               1.09   Skew:
0.02
Prob(H) (two-sided):                  0.22   Kurtosis:
11.26
==============================================================================
=========

Warnings:
[1] Covariance matrix calculated using the outer product of gradients
(complex-step).
```

Figure 14 – Auto ARIMA results for Biolase training set.

After this was completed, we need to look at the residual plots for each stock. These plots show us whether the model was a good fit. Each individual chart is interpreted in the following way:

1) Standardized Residual – determines whether the model has a uniform variance.
2) Histogram plus density – determines if there is a normal distribution of the data with a mean equal to zero.
3) Quantiles – determines the shape of the distribution. All of the dots should fall in line with the red line shown in the graph, and any significant deviations will show the distribution is skewed.
4) Correlogram – shows if there are any correlations in the residual errors. Any autocorrelation would imply that there are some errors which the model cannot account for.

The model for the Inuvo stock is shown in Figure 15. The standardized residuals show a uniform variance, except for one spike, which is possibly an outlier. The Histogram shows a normal distribution with the mean near zero. The quantiles graph shows no skewness as nearly all of the plots fall on the red line. The correlogram shows that there are no correlations in the errors.

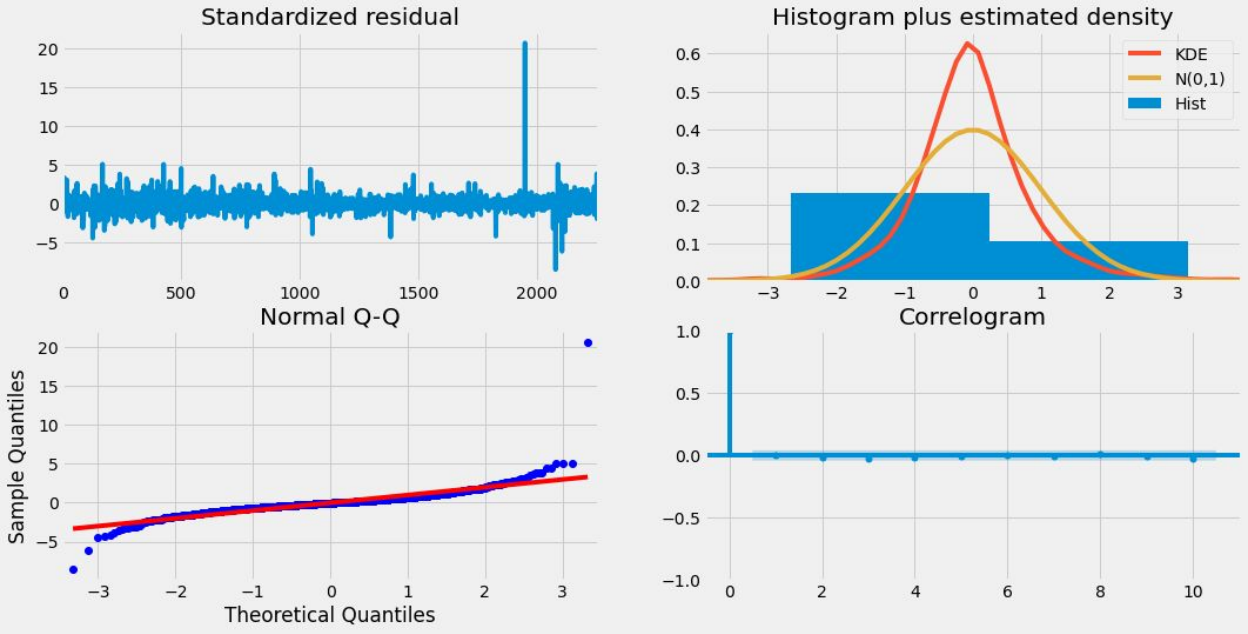

Figure 15 – Diagnostic plots for Inuvo.

The Biolase model shows a similar story. The standardized residuals show a uniform variance, with any spikes upward being offset by the same spike in the opposite direction, which implies the model has a uniform variance. The histogram plots show a normal distribution with a mean of zero. The quantiles plot shows no skewness as most of the plots fall within the red line, except at the tails. The correlogram shows that there are no correlations in the errors. See Figure 16.

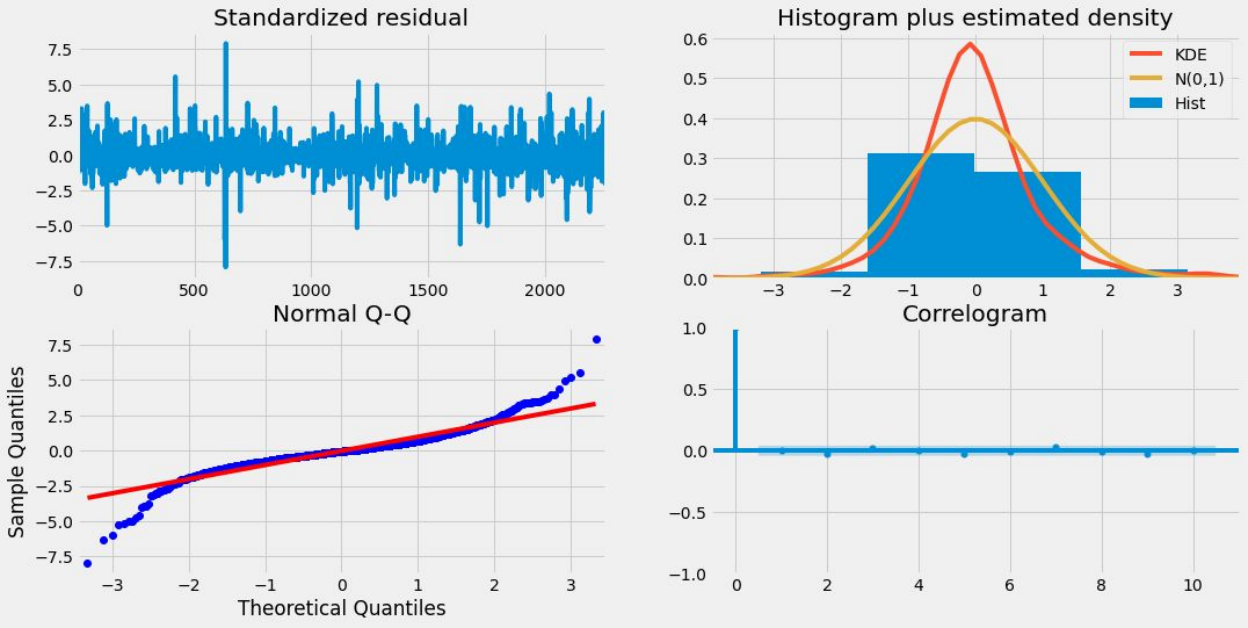

Figure 16 – Diagnostic Plots for Biolase.

After the models are fitted, we are able to run the model for predicting the price of these two stocks. The time range for this prediction is for the next year, starting at the beginning of 2020. The orange line shows the prediction of the model. The gray area is known as the cone of uncertainty, which means that the predicted price can fall anywhere around the cone. The fluctuations of the market can cause the price to go outside of the cone. The prediction for Inuvo is shown in Figure 17.

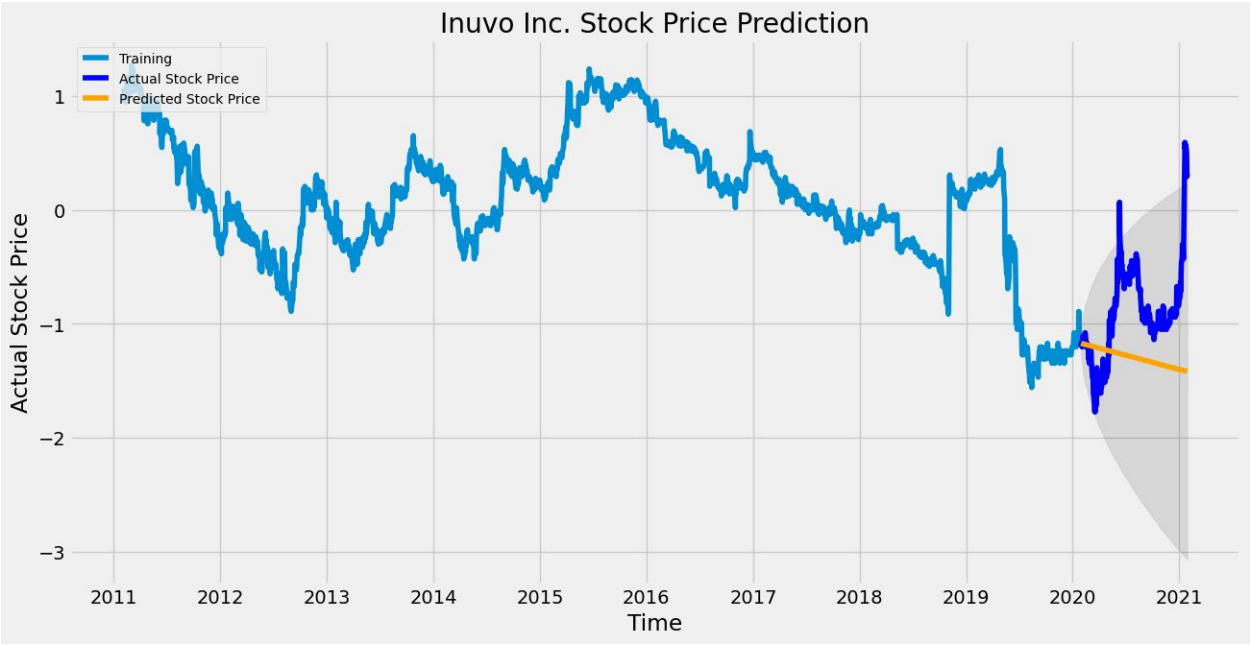

Figure 17 – Inuvo Stock price prediction.

The forecast for Inuvo is for the price of the stock to trend downwards. So far, this model is not holding because the recent value of the stock as the stock has seen some gains recently to go outside of the cone of uncertainty. When we measure the effectiveness of this model, we use the mean-squared and root-mean-squared (MSE and RMSE) to measure the error of this prediction, which is shown in the chart below.

MSE (Mean squared error)	0.3674004648899663
RMSE (Root Mean Squared Error)	0.4911668965645873

The error of this model falls way outside of the 0.05 percent confidence interval, we cannot say with confidence that the model is accurate or if the forecast will be accurate over time.

The forecast for the Biolase stock is shown in Figure 18. The model shows nearly the same prediction as the Inuvo stock, however the price of the stock, despite recent gains, the price seems to fall within the cone of uncertainty, which tells us that the model is holding true. When looking at the effectiveness of this model, while it does fall outside of the confidence interval, it is much closer than the Inuvo stock. We still cannot say with any confidence that the forecast is accurate or it will be accurate over time.

MSE (Mean Squared Error)	0.18947182182285666
RMSE (Root Mean Squared Error)	0.4352836107905473

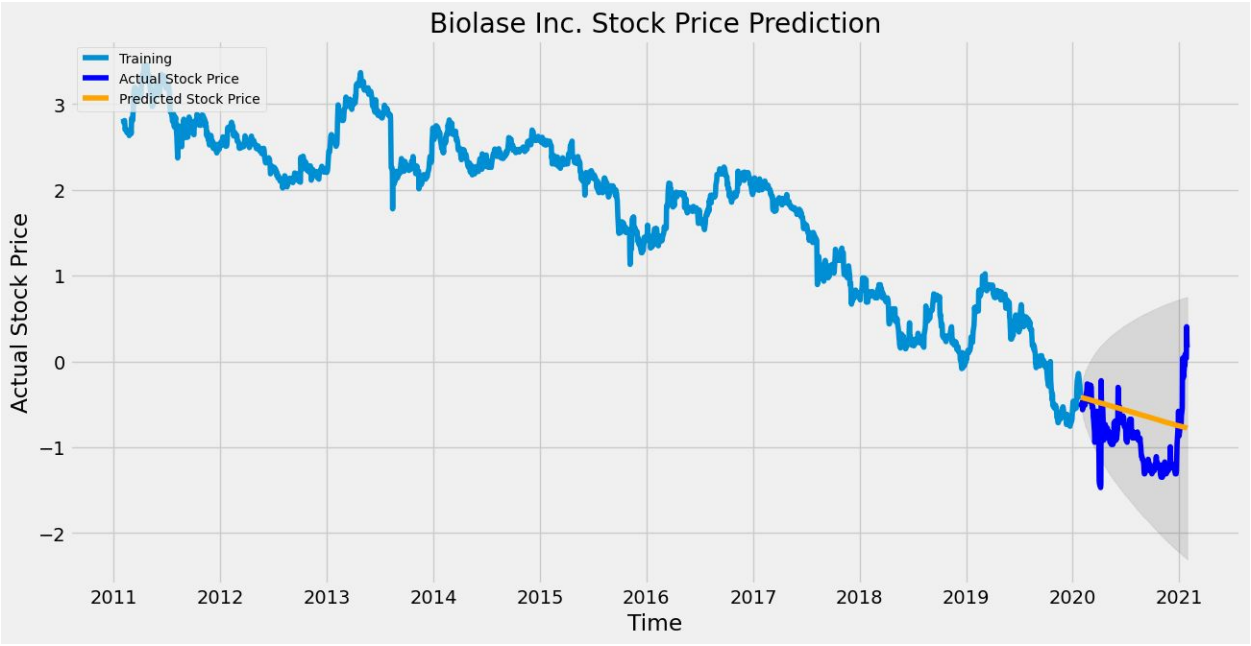

Figure 18 – Biolase Stock price prediction.

Conclusions

We can draw the following conclusions about modeling predictions with penny stocks. The prices of these stocks look very attractive to the beginner investor or anyone looking to get into the market with a minimal investment, but there is an important caveat here. Anyone looking to invest in these stocks in the short-term (e.g., day-traders looking to make a fast buck) will be disappointed by the return value of these stocks.

The model that we used to predict these stocks predicted a modest downward trend for these two stocks for the current year. However, outside forces in the market may heavily impact this prediction. The very nature of these stocks causes a remarkably high error rate when evaluating these models. Any small variation in the price of the stock will cause the error rate to fluctuate wildly, so these prediction models generated here should not be used as a tool for investing.

The accuracy of these predictions should not be trusted because of the same properties of these two stocks. Obtaining an accurate prediction of these two stocks is anyone's guess. The prices of these stocks are so low, that any monumental movement, in any direction, of these two stocks will have a tremendous impact on their prediction.