D214 Capstone- Time Series Project

Jacob Polomsky

School of Technology, College of IT, Western Governors University

Data Analytics Graduate Capstone - D214

Dr. Daniel Smith

August 25, 2024

A. To What extent can the Closing Price of SPY be predicted? This question seeks to forecast the closing price for SPY for the following year. This analysis will help traders gauge the market trend for the next year. It will also help traders anticipate where the market closing price will be over the next year so they can use it to know which trades might favor them. The context is that when trading in the stock market, 95% of traders fail due to losing money. I want to use data analysis to find patterns in the market that will result in more winning trades for our organization. We will test the result of our research question, the null hypothesis, to see if we can predict the closing price of SPY with a 90% accuracy. When we make a prediction, we have our predicted values and confidence intervals with our standard error. Our goal for the null hypothesis is to see if 90% of the predicted data can fall into at least two standard deviations from predicted values. Ideally, if 90% of the predicted values fall within one standard deviation away, that would be even better.

B. I collected the historical stock data from the last five years for SPY. In the historical data, we only need the date and the Closing price for the stock. Getting the data is relatively easy by just going to Yahoo Finance, searching for the ticker you want, in this case, SPY, then going to historical data, switching to a five-year chart, then clicking download. One advantage to this is the data is relatively quick and easy to get, allowing you to analyze multiple stocks seamlessly. One disadvantage is that the data is static, so you can't get the newest information. You would have to re-download the data to get the latest stock information, and depending on the analysis, you might need to be updated by the time the study is

C. The packages that will be used are forecasting, which is used to pick the model and evaluate the model performance. Astsa needs help checking the accuracy. However, it makes plotting easy, and the forecast is like the opposite, making checking the model's accuracy very easy, but plotting is more complicated. Tseries is to run the ad fuller test.

A screenshot of a computer

Description automatically generated

The process to prepare is to clean the data to make sure the data is in the correct format and that there are no missing values or duplicates. The step we took were loading the data, checking for null values, checking for duplicates, checking for outliers. This allows us to format the data for analysis, making it usable for the Arima or Sarma models that would not work if not formatted correctly or might throw off the analysis if not cleaned. One problem that might arise with data cleaning is that it might open up some data gaps which also might lead to the model throwing an error or throwing off the model entirely.

First was to load in the data



Next, I checked for missing values, only looking at Date and Closing. This removes missing values from the data and might open up a gap in the data.





Next was to check for duplicates. It removes duplicate values but might add a gap in the data.



Next is a summary to check for any outliers in the data. You might see outliers that throw off data, introduce gaps, or give you fewer data if you want to remove the extreme.

Then converts the data into a time series. Allows for plotting the data in Arima or Sarima model, but misses on seasonal trends with a frequency of 1 instead of dates.A close up of a sign

Description automatically generated

Then, we split the data into train and test sets to test how well our model performs. This allows us to train the model and predict values the model has not seen before, allowing you to understand better how the model will perform on real-world datasets. A disadvantage is that it leaves you with less data, making the model less accurate than if you left the data in.A black text on a white background

Description automatically generated

D.



A screenshot of a graph

Description automatically generated

This model shows a tailing off ACF, and the PACF cuts off, indicating that it is an AR model.

For finding if the data is stationary for d, with the plot below, we find that the p-value is above 0.05, indicating that this data is non-stationary.



A white background with black text

Description automatically generated

Checking for seasonal components we run a decomposition plot

A close up of a text

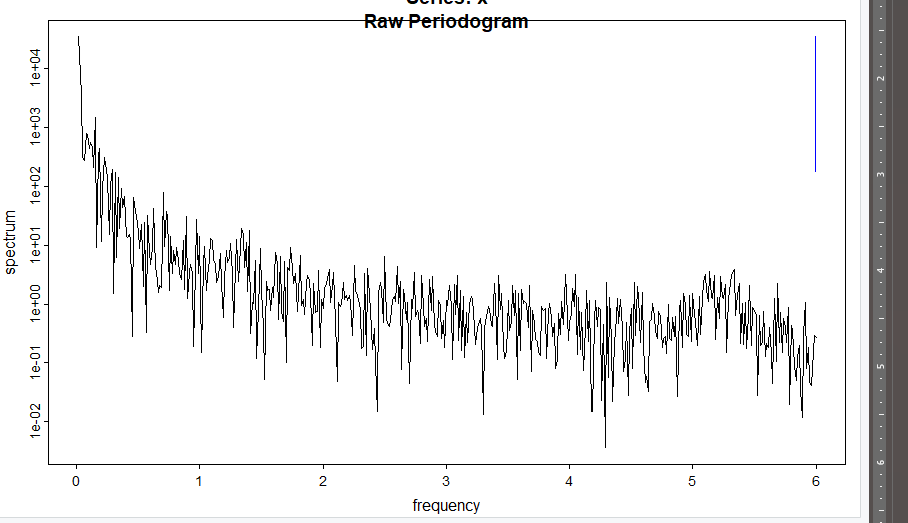
Description automatically generated

A graph of different types of waves

Description automatically generated with medium confidence

We see an increasing trend over time, but towards the end of the plot, it starts to trail down; the seasonal component is very repetitive, indicating a seasonal correlation with the plot.

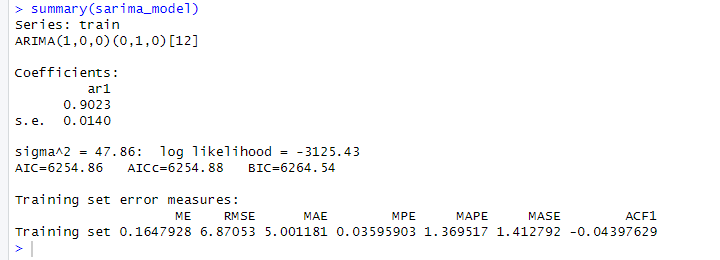
 We use the spectrum plot to see if the data is stationary or not stationary.



Model 1 is first choosing a 1 0 0 0 1 0 12 model because we have an AR model due to the ACF and PACF plot, and D of 1 because of the seasonality of the decomposition plot.A close up of text

Description automatically generated







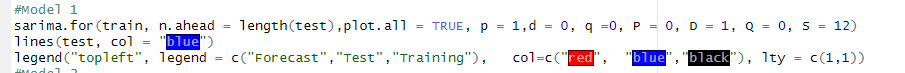
A screenshot of a graph

Description automatically generated



A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

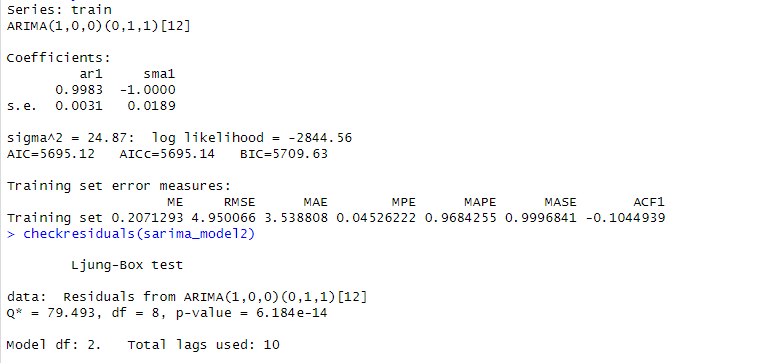


A graph with a line and a circle

Description automatically generated with medium confidence

Model 2  I did the same thing as model one but noticed that the residual for ACF had a big residual at 12, showing that there might be a missing seasonal component, so I added a Q of 1A white background with black text

Description automatically generated

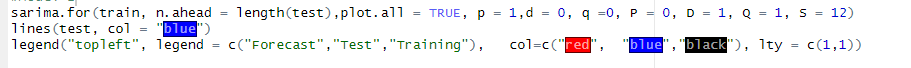


A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated



A graph with a line pointing to the top

Description automatically generated with medium confidence

For the first model, I went with a 1 0 0 0 1 0 12 model because of the ACF trailing off and PACF cutting off, giving a p of 1 and a q of 0. I went with a seasonal D of 1 instead of d because, in the decomposed plot, the chart showed a seasonal pattern. I plotted it with both D and d, but it threw off the analysis. With the second model, I noticed that Lag 12 had a big stick when ACF residual indicated a possible seasonal component we needed to include. So I did a 1 0 0 0 1 1 12 model, and it looks like the model performed better. One advantage is that the model follows the data relatively well and gets the direction the market is going with pretty good accuracy. One disadvantage is that the model is missing from the overall trend.

E. With these models, we were able to get a MAPE of 1.3 for the first model and 0.9 for the second model, showing that we could predict the data very well and that we could predict over 90% of the data, which is our goal. The MAE was five, and the RMSE was 6 for the first model, showing that the model was off by 5 or 6. The predictions are close when you factor in variables around 300 and 500. The second model had an MAE of 3 and an RMSE of 4, showing that this model performed better than the first model by almost every metric: RMSE, MAPE, MAE, AIC, and BIC. When looking at the plot, you can see that the plot for model 1 has a wide margin of error, with model 2 having a lower margin of error, but the data falls into a higher confidence ratio when you plot the test set. With models, they predict closing prices well, but it still needs to be fixed. It will be helpful for traders to know where to expect the market to be a year from now, but I don't know how well the model would be used to understand which trades to make. I recommend trying a more advanced model to narrow the margins so that trades are more profitable if you want to use this to make trades. I would recommend one of two things first is maybe to do another stock that might find the data better for a time series analysis and trade that stock, or maybe to switch this to a hybrid model or another machine learning model that might be able to trade volume and starting price to get more accurate predictions.