Final Project:
Forecasting Amazon Net Sales by Using ARMA and ARIMA Models

Authors: Levi Lan, Xingyuan Ding, Yu-Hsing Chang
Professor: David Rios
Date: 12/16/2022

1. Introduction

Our team uses the time series model to fit and forecast the Amazon Net sales. We divide our report into 4 sections. In Section 2, we plot the Amazon Net sales to see how this time series looks like. And we do the Dickey-Fuller test to check its stationarity. After that, if the data is non stationary, we will detrend and deseason our data to get the residual dataset for model use. Secondly, we do the stationary test on residual dataset, and plot the ACF and PACF to take a guess on our model parameters. In this project, we decided to use ARMA and ARIMA models, which in Section 3 and 4 respectively, to forecast the data. Section 5 discusses the future improvement and we will make a brief summary of our findings at the end.

2. Data

2-1 Data Description
In this project, we use Amazon quarterly net sales from 2000 Q1 to 2022 Q3 as our original data. This total value of all net sales includes net product sales and service sales, and it also combines North America sales, international sales, and AWS (Amazon Web Services) sales. Let's draw the figure first to see how the data looks like.

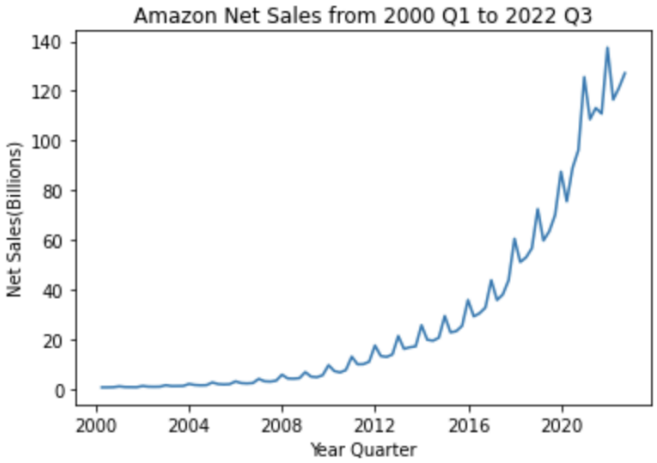


Figure 1. Amazon Net Sales from 2000 Q1 to 2019 Q3

The figure above shows that there is an obvious trend and seasonality of the Amazon quarterly net sales, which means that it could be a proper times series data for the purpose of our project. We now have 91 data points in total, the second thing is to split the data into

training data and testing data. We use data before 2019 as our training data, and the remaining data as our testing data.

Next, we use the Augmented Dickey-Fuller test to check the stationary of our training data. Augmented Dickey-Fuller test is a common statistical test used to test whether a given time series is stationary or not. We will reject the null hypothesis if the time series is stationary, otherwise we cannot reject the null hypothesis. Here, we set the confidence level of 95% as the threshold. The results of the ADF test describe that the p-value is greater than 0.05; therefore, the training data is not stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                  2.571972
p-value                         0.999069
#Lags Used                     12.000000
Number of Observations Used    63.000000
Critical Value (1%)            -3.538695
Critical Value (5%)            -2.908645
Critical Value (10%)           -2.591897
dtype: float64
```

Figure 2. Stationary Test on Training Data

2-2 Detrend and Deseasonalized

One way to detrend time series data is to fit a regression model to the data. In this project, we apply linear regression as our model. From the above graph, our data grows exponentially; therefore, we first log the data to make it fit to our linear regression model. The figure following shows that the log of training data looks linear.

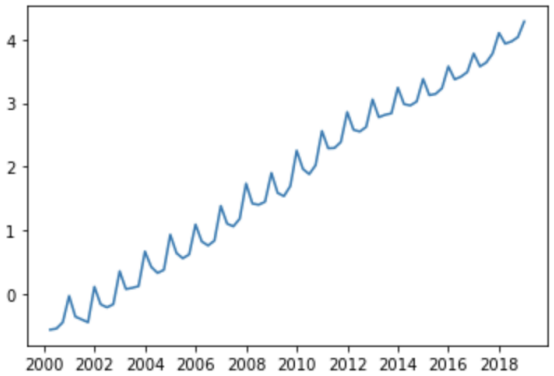


Figure 3. Log of Training Data

After transforming the data, we assign from 0 to 62 as our timestep and fit it to the observed log of net sales. We can calculate the residual which represents the detrended data by differencing between the actual value and the predicted value for each observation.

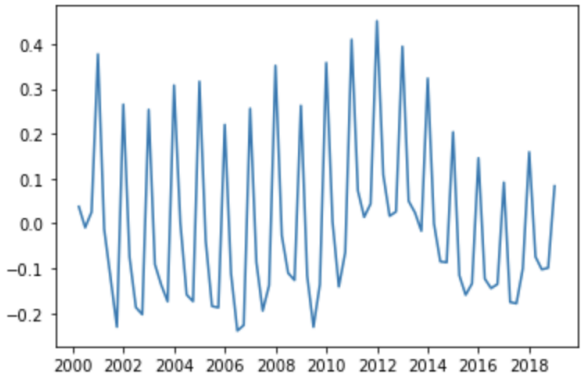


Figure 4. Detrended Training Data

So far, we have already detrend the log of training data. Next, we apply the package called seasonal_decompose in python to show the seasonality, and then the difference between the detrended data and the seasonality would be our residual dataset.

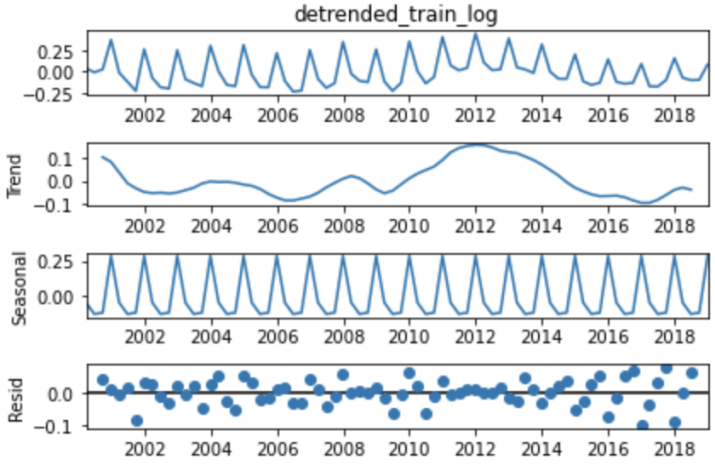


Figure 5. Decomposition of Detrended Training Data

3. ARMA model

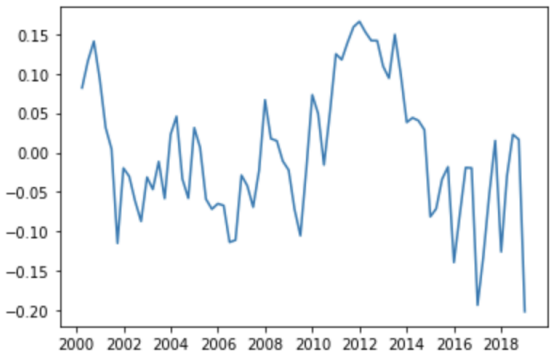
3-1 Introduction to ARMA

ARMA(p, q) is a forecasting model that applies autoregressive analysis (AR) and moving average (MA) methods to well-behaved time series data. ARMA assumes that the time series is stationary, and if it fluctuates, it fluctuates uniformly at specific points in time. The moving-average model specifies that the output variable is cross-correlated with a non-identical to itself random-variable. An autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term.

3-2 Forecasting

Now, we will use the residual dataset which is already detrended and deseasonalized to do the forecasting. Since the ARMA model has an assumption of stationary time series, we test the

stationarity of the residual dataset. The p-value decreases a lot compared to the original data but is still greater than 0.05, meaning that our residual dataset is non-stationary. Even though we know that this dataset is not appropriate for the ARMA model, we still want to complete our model to verify whether the result is worse or not.




```
Results of Dickey-Fuller Test:
Test Statistic                  -2.236161
p-value                          0.193384
#Lags Used                       8.000000
Number of Observations Used     67.000000
Critical Value (1%)             -3.531955
Critical Value (5%)             -2.905755
Critical Value (10%)            -2.590357
dtype: float64
```

Figure 6. Stationary Test on Residual Dataset

To begin with the ARMA(p, q) model, the first thing is to plot the PACF and ACF graphs to take a guess on p and q respectively. The Partial Autocorrelation Function (PACF) is a conditional correlation that shows the partial correlation between a stationary time series and its own lag values. In addition, PACF plays an important role in identifying lag of p in AR models. Autocorrelation Function is the correlation of a signal with a delayed copy of itself as a function of delay, and it also can estimate the value of q in MA model. We can clearly see the tail off in both ACF and PACF graphs, meaning that ARMA is a proper model for forecasting. In this situation, we select q=4 based on the ACF and select p=5 based on the PACF. Overall, our model would be the ARMA(5,4).

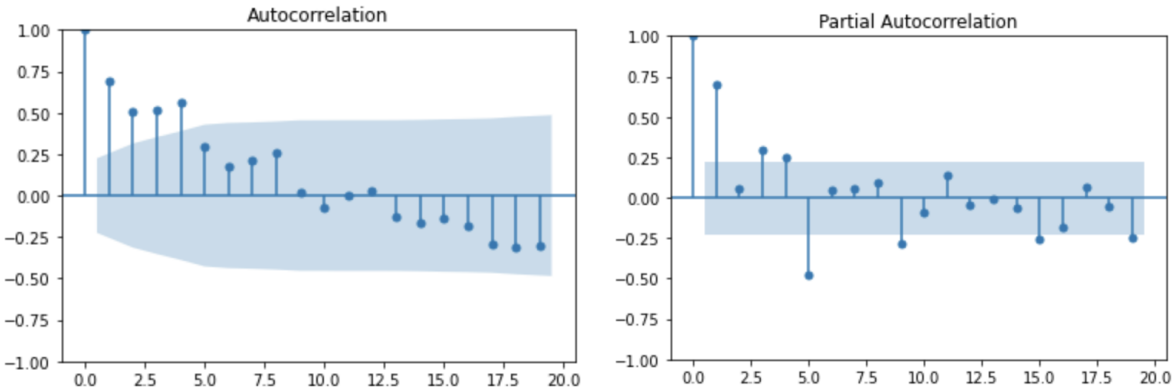


Figure 7. ACF and PACF

After fitting the ARMA(5,4) model, we forecast the testing data to see its performance. The following figure shows that the forecasting data doesn't fit really well to the actual testing values. The reason may be that the residual dataset we use is not stationary or the recession in the end of 2022 makes the actual net sales grow more slowly than before. To deal with the first explanation, we will apply ARIMA as our second model, and try to improve the model performance.

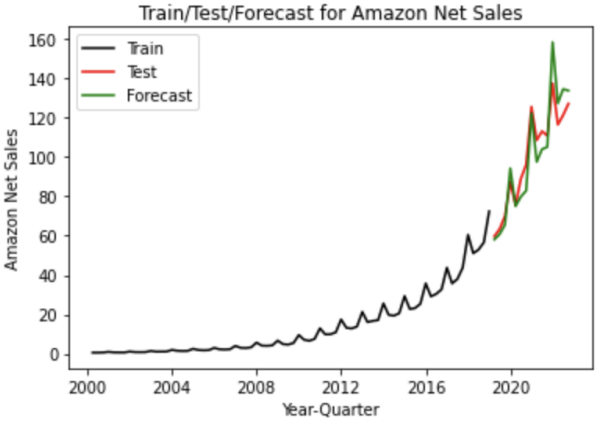


Figure 8. Train, Test, and Forecast for Amazon Net Sales

4. ARIMA Model

4-1 Introduction to ARIMA

As a potential improvement, we would like to introduce the autoregressive integrated moving average (ARIMA) model. An ARIMA model is a generalization of an autoregressive moving average (ARMA) model. ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend).

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

An ARIMA model is characterized by 3 terms: p, d, q. p is the order of the AR term. q is the order of the MA term. d is the number of differencing required to make the time series stationary.

4-2 Parameter Selection

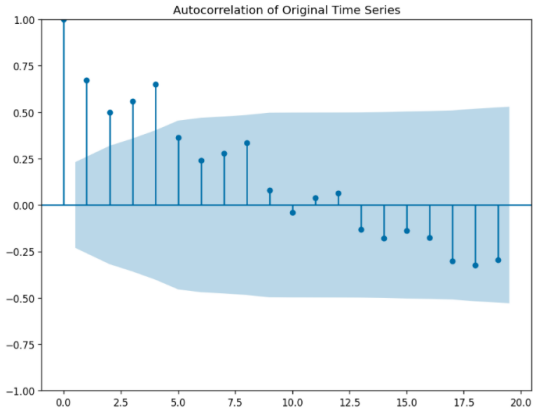
First step is to check if the series is non-stationary. Else, no differencing is needed, that is, d=0. The data conducting the Augmented Dickey-Fuller Test is time series after detrending

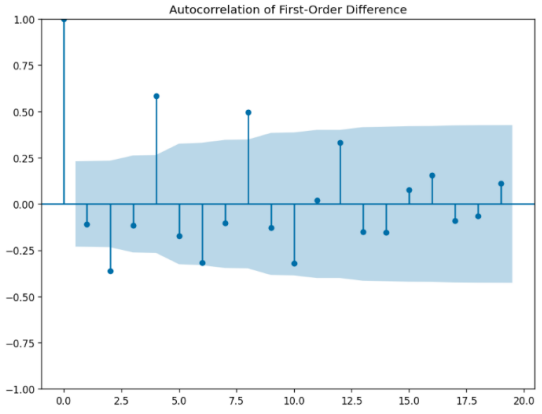
and de-seasoning. The p-value is 0.23, which is larger than 0.05. It fails to reject the null hypothesis and infers that the time series is non-stationary. Hence, we could proceed to build the ARIMA model.

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.117010
p-value                          0.237730
#Lags Used                       8.000000
Number of Observations Used     63.000000
Critical Value (1%)             -3.538695
Critical Value (5%)             -2.908645
Critical Value (10%)            -2.591897
dtype: float64
```

Figure 9. Augmented Dickey-Fuller Test for the time series after detrending and de-seasoning

First step is to select the order of differencing (d). We differentiate the series and plot autocorrelation charts. For the original series, we could find the autocorrelations are positive for many numbers of lags (1 to 7). It could be another evidence that the original time series is not stationary. For the other two series, the time series reaches stationarity with the first order of differencing. Second-order difference does not improve much. So, we tentatively fix the order of differencing as 1 even though the series is not perfectly stationary (weak stationarity).





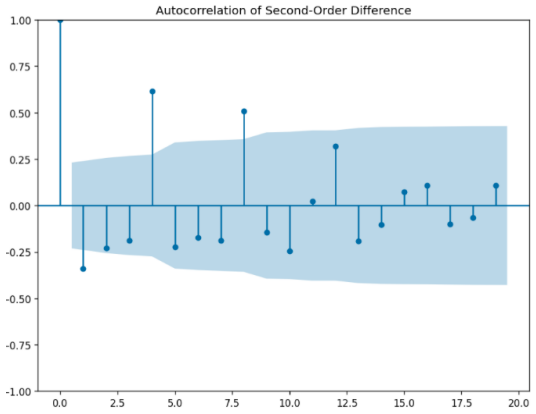
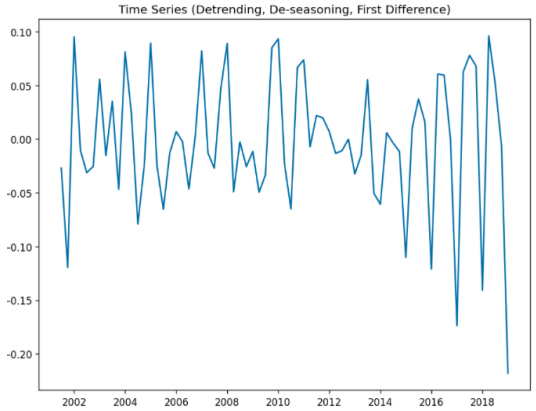


Figure 10. Autocorrelation Charts of Original Series, First-Order Difference, Second-Order Difference

The time series after taking the first difference shows below. Then,we conduct the Augmented Dickey-Fuller Test again. The p-value is 0.007, which is less than 0.05. So, It rejects the null hypothesis and infers that the time series is stationary.




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.523741
p-value                          0.007398
#Lags Used                       4.000000
Number of Observations Used     66.000000
Critical Value (1%)             -3.533560
Critical Value (5%)             -2.906444
Critical Value (10%)            -2.590724
dtype: float64
```

Figure 11. Time Series After Taking First Difference and Its Augmented Dickey-Fuller Test

Next, we would determine the order of AR Term (p). According to the (Partial Autocorrelation) PACF plot below, the PACF lag 4 is quite significant since it is well above the significance line(blue region). So, we tentatively fix the p as 4.

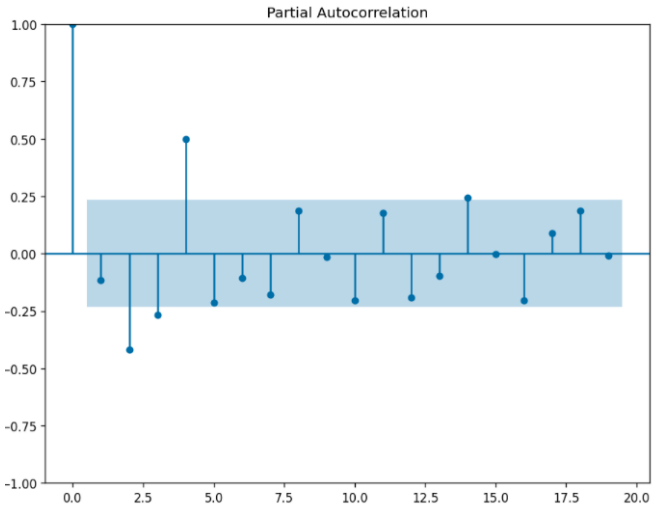


Figure 12. Partial Autocorrelation for Time Series After Taking First Difference

Lastly, just like how we looked at the PACF plot for the number of AR terms (p), we can look at the ACF plot for the number of MA terms (q). As can be observed in figure above, either 4 or 8 could be good candidates for q.

4-3 ARIMA Model

We try both ARIMA(4,1,4) and ARIMA(4,1,8) and compare them with the Akaike Information Criterion (AIC). ARIMA(4,1,4) has AIC -233.42, and ARIMA(4,1,8) has AIC -233.67. Therefore, ARIMA(4,1,8) is a better model. Then, we check the residuals. There is no specific pattern, and the density, which is similar to the normal distribution also looks okay.

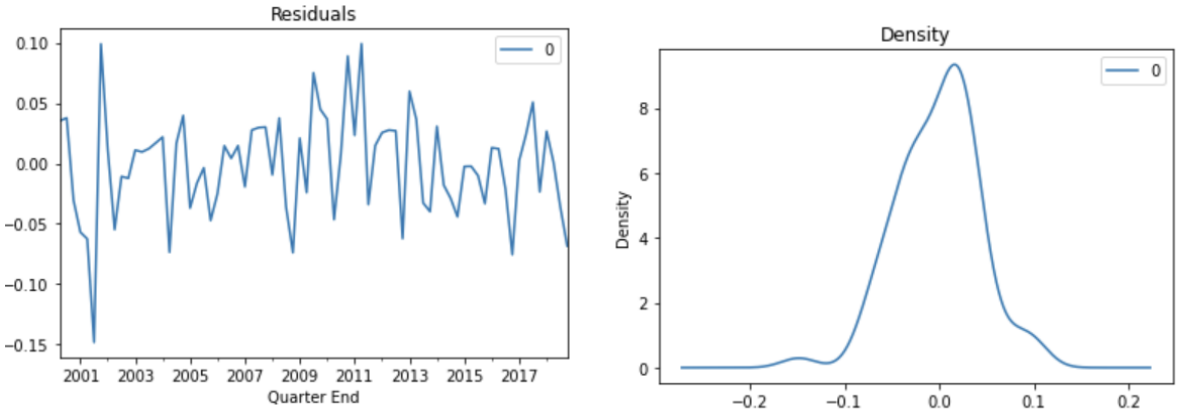


Figure 13. ARIMA(4,1,8) Model Residuals

4-4 Prediction

Now, we use ARIMA(4,1,8) to make predictions on testing data. The visualized result is shown below. Overall, the model captures the general movement of net sales. The predictions and testing data are close though sometimes a little bit higher or lower. Mean Square error of prediction is 87.41.

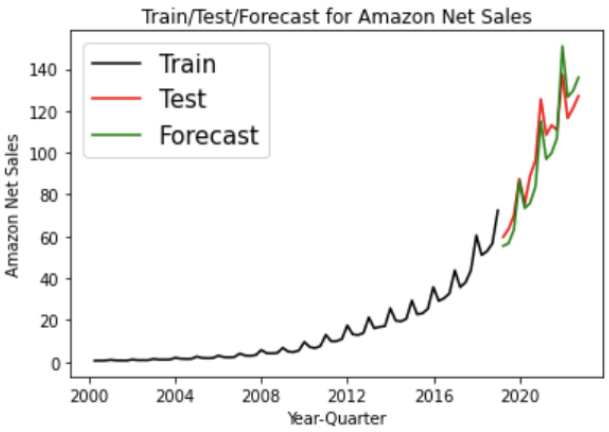


Figure 14. Train/ Test/ Forecast for Amazon Net Sales

5. Conclusion & Further Improvement

In this project, we use data before 2019 to train the ARMA and ARIMA and make predictions on 2020. Before putting into the ARMA model, we take log and make the time series detrended and de-seasoned. According to ACF and PACF, we get our parameter estimates as ARMA(5,4). The RMSE on testing data is 9.52. However, the time series after detrending and de-seasoning are still non-stationary. Thus, ARIMA is introduced. Similar to ARMA, we estimate our parameters by ACF and PACF and get the best set of ARIMA(4,1,8). The RMSE is 9.35, relatively better than the ARMA model.

A potential problem with our approach is that we assume exponential growth of Amazon net sales, which is unlikely to be the case in the longer term. Hence, we should revisit our model periodically and retrain it with a fixed lookback window rather than a fixed starting year. If a slowdown in Amazon net sales growth is detected, we may even need to update the detrending methodology from taking the logarithms of the sales to square roots or something similar.