Project: Forecasting Sales

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/edd0e8e8-158f-4044-9468-3e08fd08cbf8/project

Step 1: Plan Your Analysis

Look at your data set and determine whether the data is appropriate to use time series models. Determine which records should be held for validation later on (250 word limit).

Answer the following questions to help you plan out your analysis:

1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.
2. Which records should be used as the holdout sample?

The dataset meets the criteria of a time series dataset, since:

1. It is over a continuous time interval from 2008-01 to 2013-09;
2. There are sequential measurements (every month sales) across that interval;
3. There is equal spacing (every month) between every two consecutive measurements;
4. Each time unit within the time interval has at most one data point.

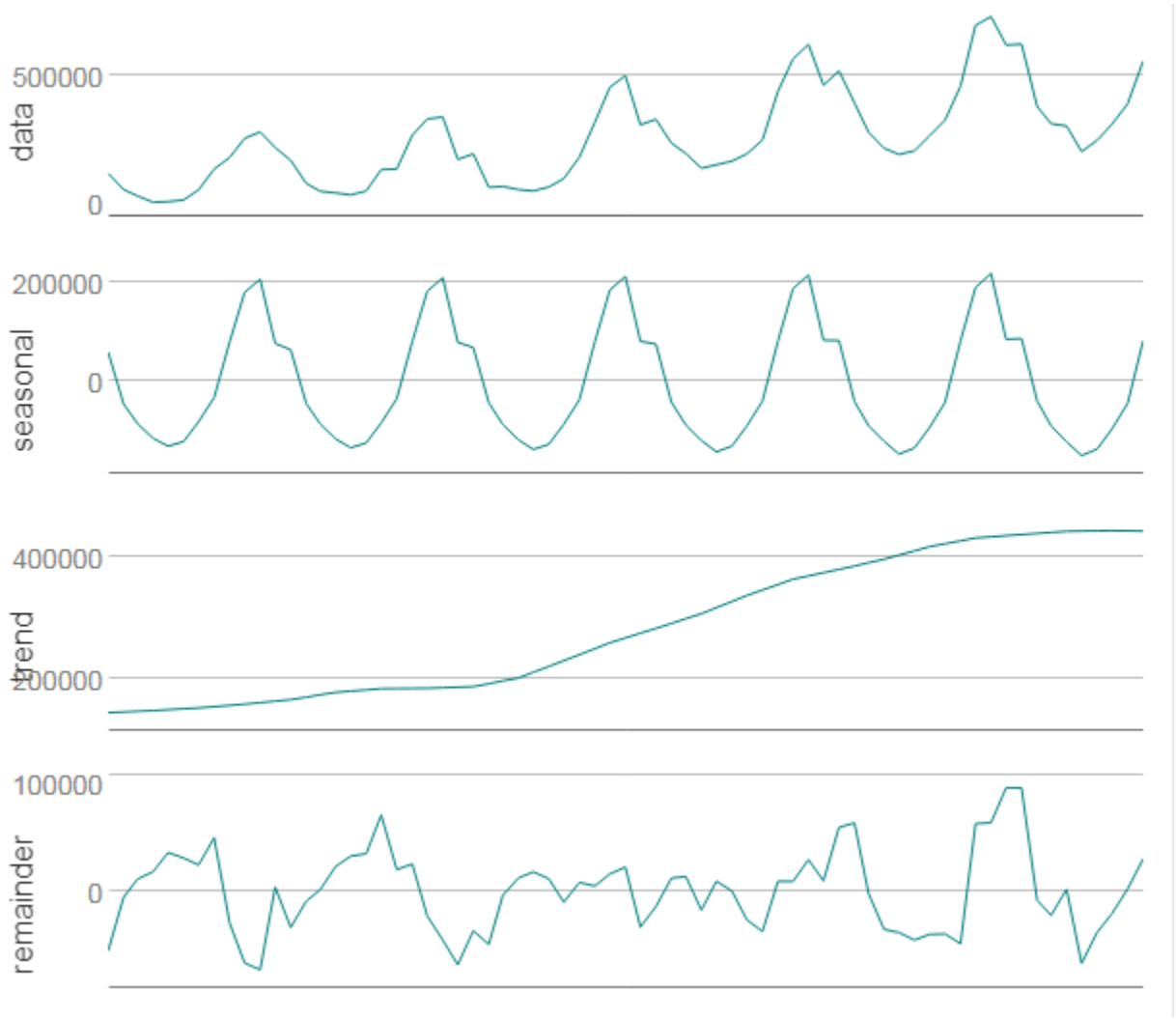
Since I need to forecast the next 4 months of sales, the holdout sample size should be at least 4 months and be the most recent data. Since we have 69 measurements, I will filter out the last 4 months as holdout and use ID equal or less 65 to do forecast. And leave the 4 months of 2013-06 to 2013-09 for holdout sample.

Step 2: Determine Trend, Seasonal, and Error components

Graph the data set and decompose the time series into its three main components: trend, seasonality, and error. *(250 word limit)*

Answer this question:

1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.



By viewing this time series decomposition plot, we can see there is a constant seasonal pattern, a linear pattern of the trend line, and the error showing changing variance.
Therefore, the relatively constant seasonal variation should apply the additive method. The additive method applies to the trend, since its variation increase in magnitude overtime (the exponential behavior). And the error should be used multiplicatively.

But I also tried the auto set of the ETS composition, and received a M,A,M recommendation. So I reviewed the graph of season again and hover the mouse over each peak to check the number and I found that there is an increase. Therefore, the season part should be considered M.
I decided to try both M, A, A and M, A, M for safe.

Step 3: Build your Models

Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)

Answer these questions:

1. What are the model terms for ETS? Explain why you chose those terms.
    a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

I tried both ETS(M,A,A) and ETS(M,A_DAMPED,A), since the error is increasing and shall be used multiplicative, the trend is linear and shall be used additively, and the seasonal pattern is constant and shall be used additively.

At the same time, I tried the recommended set of ETS(M, A, M).

For the ETS (M,A,A), the RMSE is 48206.68, the AIC is 1673.4, and the MASE is 0.53. For the ETS (M,A_D,A), the RMSE is 36989.32, the AIC is 1690.5, and the MASE is 0.39. For the ETS(M, A, M), the RMSE is 33153.53, the AIC is 1639.5, and the MASE is 0.37.

For the ETS (M,A,A), the in-sample error measures are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-3165.523166	48206.6800824	36215.8613811	-4.3453583	14.7897347	0.5283349	0.6440806

For the ETS(M, A, M), the in-sample error measures are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

For the ETS (M,A_D,A), the in-sample error measures are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-29.1215006	36989.3153041	26518.3053615	0.9565698	12.4111002	0.3868621	0.0483172

Two key components to look at are RMSE and MASE. The RMSE is the sample deviation of the differences between predicted values and observed values. The MASE is the mean absolute error of the model divided by the mean absolute value of the first difference of the series, and can be used to compare forecasts of different models.

Since the less the RMSE, the better, we can see the ETS(M,A,M) has a least RMSE of 33153. And the less the MASE, the better, we can see the ETS(M,A,M) has a least MASE of 0.36, a value falling below the generic 1.0, the commonly accepted MASE threshold for model accuracy.

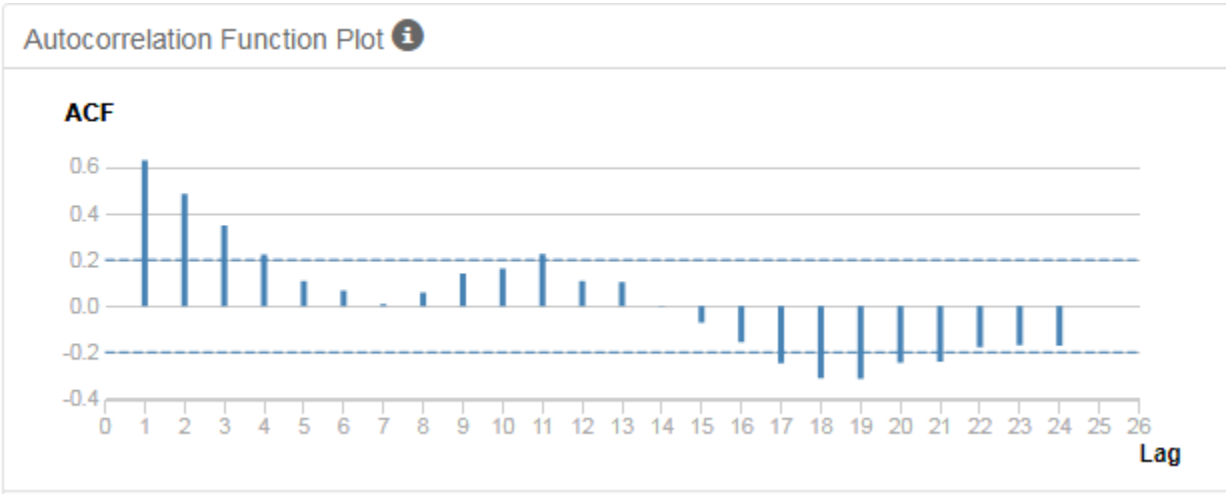
Compared three possible models, the ETS(M, A, M) is the best.

2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for

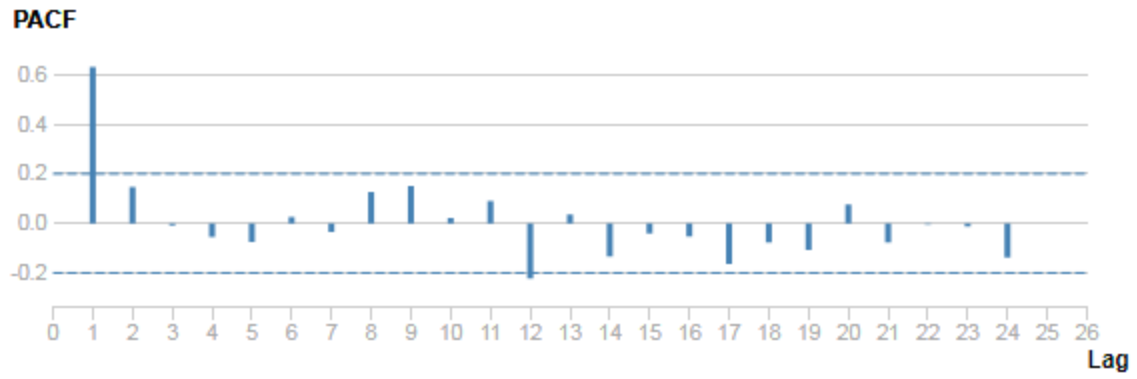
the time series and seasonal component and use these graphs to justify choosing your model terms.

a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

The graph after first differencing are:

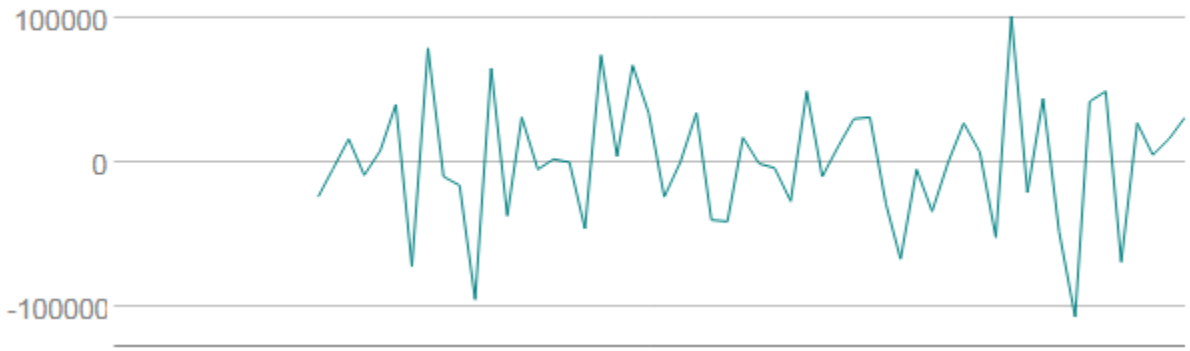


Partial Autocorrelation Function Plot

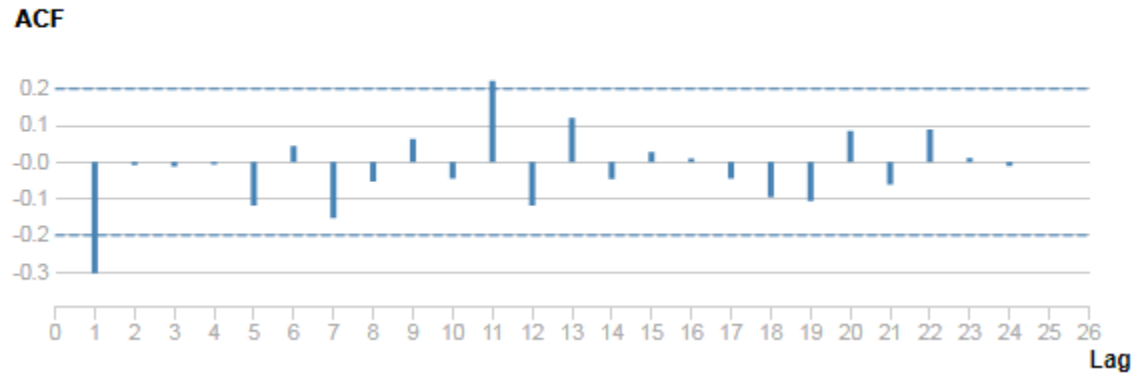


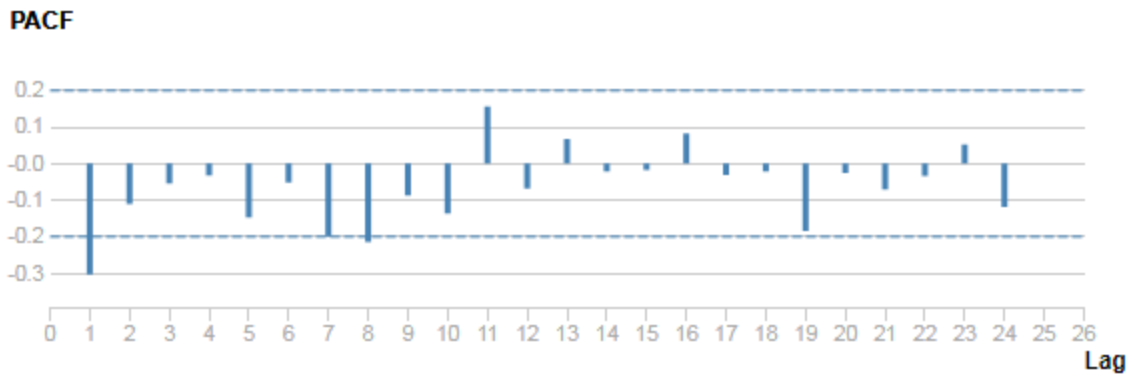
The graph after another difference (the seasonal differencing) are:

Time Series Plot



Autocorrelation Function Plot





Based on the two time series graphs, we could see that after one integrated differencing and one seasonal differencing, the time series is stationary. The current ACF lag-1 is negative. The ACF displays a sharp cutoff. The PACF gradually decreases. Therefore, it is a MA model.

Further, the ACF cuts off at lag-1 indicating MA number is 1. And since we used one differencing to make the time series stationary, the I is 1. The ARIMA model by now is ARIMA(0,1,1).

Since the pattern is seasonal, we also need to consider the seasonal part of the ARIMA, the P, D, and Q. I have already done one seasonal differencing, so the D is 1. And the number of periods in each season is 12 for the year-based model. Therefore m is 12.

The ACF plot shows a negative seasonal autocorrelation in lag-1. The spikes in the PACF decays to zero, while the ACF cuts off to zero. So the seasonal MA model is appropriate. The ACF does not have a significant autocorrelation at the first seasonal period like 12. So I will try Q=0.

Therefore, the current seasonal ARIMA would be ARIMA(0,1,1)(0,1,0)(12). But since the model is clearly a MA model, but the MA Q is not so clear, I still tried another set with a Q=1, to compare the results.

For ARIMA(0,1,1)(0,1,0)(12), the in-sample errors are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

For ARIMA(0,1,1)(0,1,1)(12), the in-sample errors are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-358.1274828	36758.4027043	24996.5435416	-1.800917	9.8272386	0.3646619	0.0166958

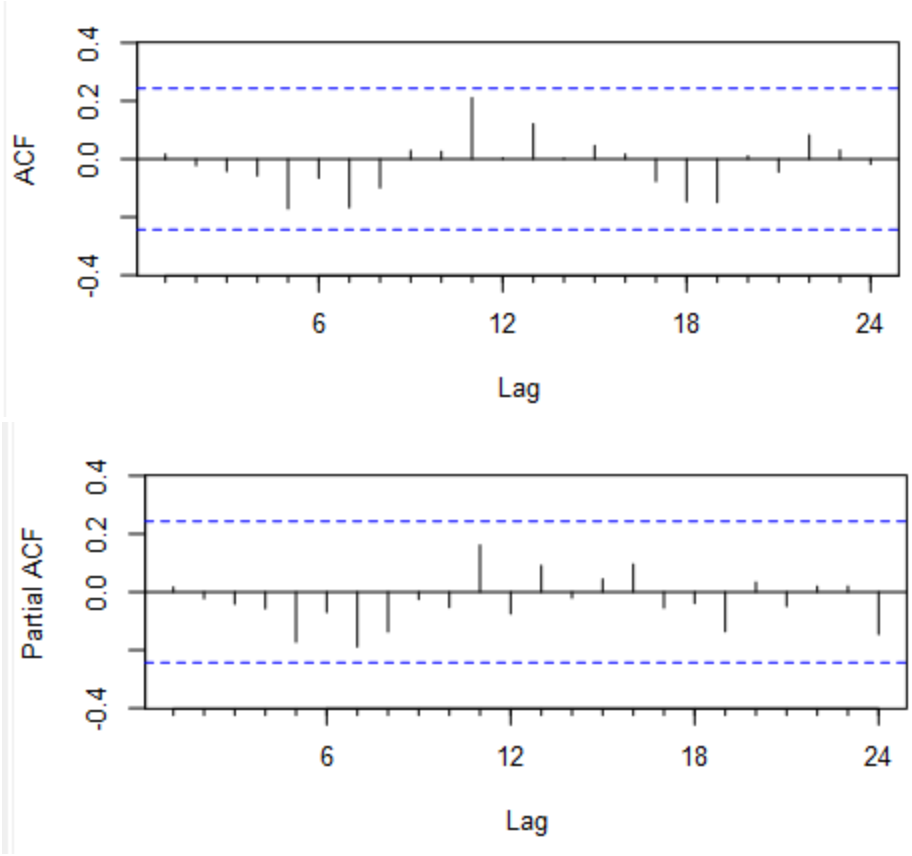
Two key components to look at are RMSE and MASE. The RMSE is the sample deviation of the differences between predicted values and observed values. The MASE is the mean

absolute error of the model divided by the mean absolute value of the first difference of the series, and can be used to compare forecasts of different models.

Since the less the RMSE, the better; the less the AIC, the better; the less the MASE, the better; , for ARIMA(0,1,1)(0,1,0)(12), AIC is 1256.5967; RMSE is 36761.5281724 and MASE is 0.3646109; for ARIMA(0,1,1)(0,1,1)(12), AIC is 1258.5932; RMSE is 36758.4027043 and MASE is 0.3646619. The auto set confirmed ARIMA(0,1,1)(0,1,0)(12).

We can see the two sets are very close. ARIMA(0,1,1)(0,1,0)(12) is better performed in AIC, but not as good as ARIMA(0,1,1)(0,1,1)(12) in RMSE. I choose ARIMA(0,1,1)(0,1,0)(12) since AIC is more important.

To regraph ACF and PACF for both the Time Series and Seasonal Difference of ARIMA(0,1,1)(0,1,0)(12), we have:



Step 4: Forecast

Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)

Answer these questions.

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

For the EST(M, A, M), the forecast error measurements against the holdout sample are:

Actual and Forecast Values:

Actual	MAM
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MAM	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA

For the EST(M, A, M), the in-sample error measurements are:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

For ARIMA(0,1,1)(0,1,0)(12), the forecast error measurements against the holdout sample are:

Actual and Forecast Values:

Actual	ARIMA_0_1_1__0_1_0_12
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_0_1_1__0_1_0_12	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

For ARIMA(0,1,1)(0,1,0)(12), the in-sample error measurements:

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

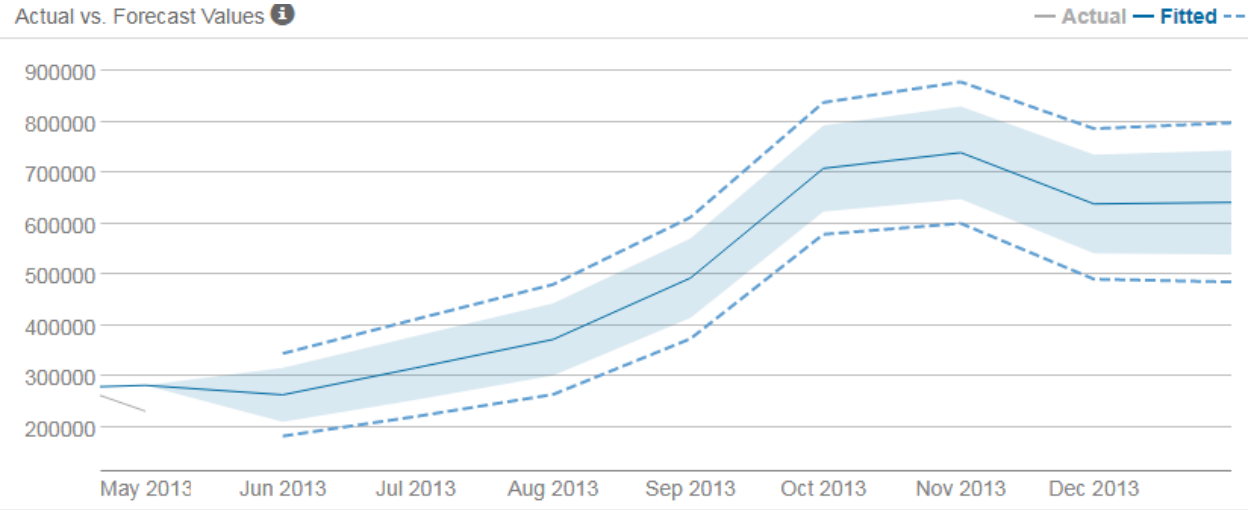
Based upon the two sets of in-sample error measurements, the RMSE and MASE of the ARIMA is lower than the ETS, meaning the deviation of the differences between predicted values and observed values of the ARIMA is less than the ETS.

Based upon the forecast error measurements against the holdout sample, the MAPE and ME of the ARIMA are lower than the ETS, meaning the average of the difference between actual and forecasted values of ARIMA is less than ETS.
Therefore, the ARIMA model is better. I choose the ARIMA model.

The forecast for the next four periods are:

Period	Sub_Period	forecast_ARIMA_0_1_1__0_1_0_12
2013	10	709228.480132
2013	11	740228.480132
2013	12	639228.480132
2014	1	642228.480132

The graph using 95% and 80% confidence intervals is:



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.