Project: Forecasting Sales

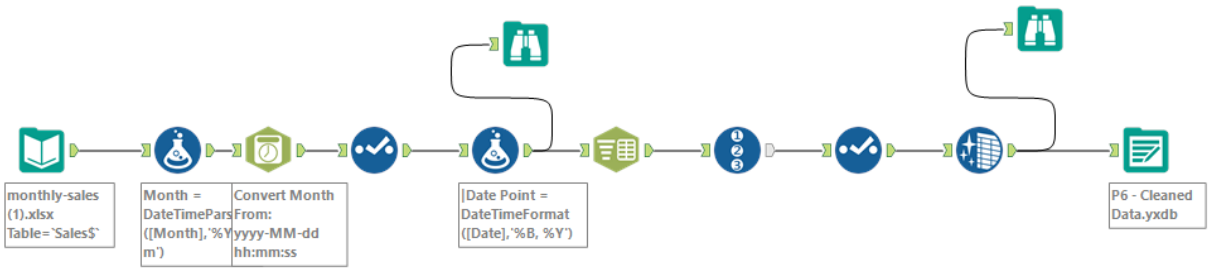
Step 1: Plan Your Analysis

To be considered a time series dataset, our data should have the following four characteristics (as outlined in Lesson 1.2: Introduction to Time Series):

1. Data should span over a continuous time interval without any missing data within this interval.
2. There should be sequential measurements across the interval of our time series.
3. Equal spacing is required between every two consecutive measurements.
4. Each time unit within the time interval has one data point at most.

Investigating the data, we see that the first entry is January 2008 and the last one is September 2013, which is 69 months in total. This is exactly as many records as there are in our data set. Transforming, the month into a Year column and a Month column we can doublecheck that indeed all the above requirements are satisfied. Hence, this is a time series.

Alteryx workflow and results format:



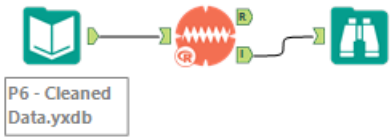
4 of 4 Fields | Cell Viewer | 69 records displayed, 2975 bytes

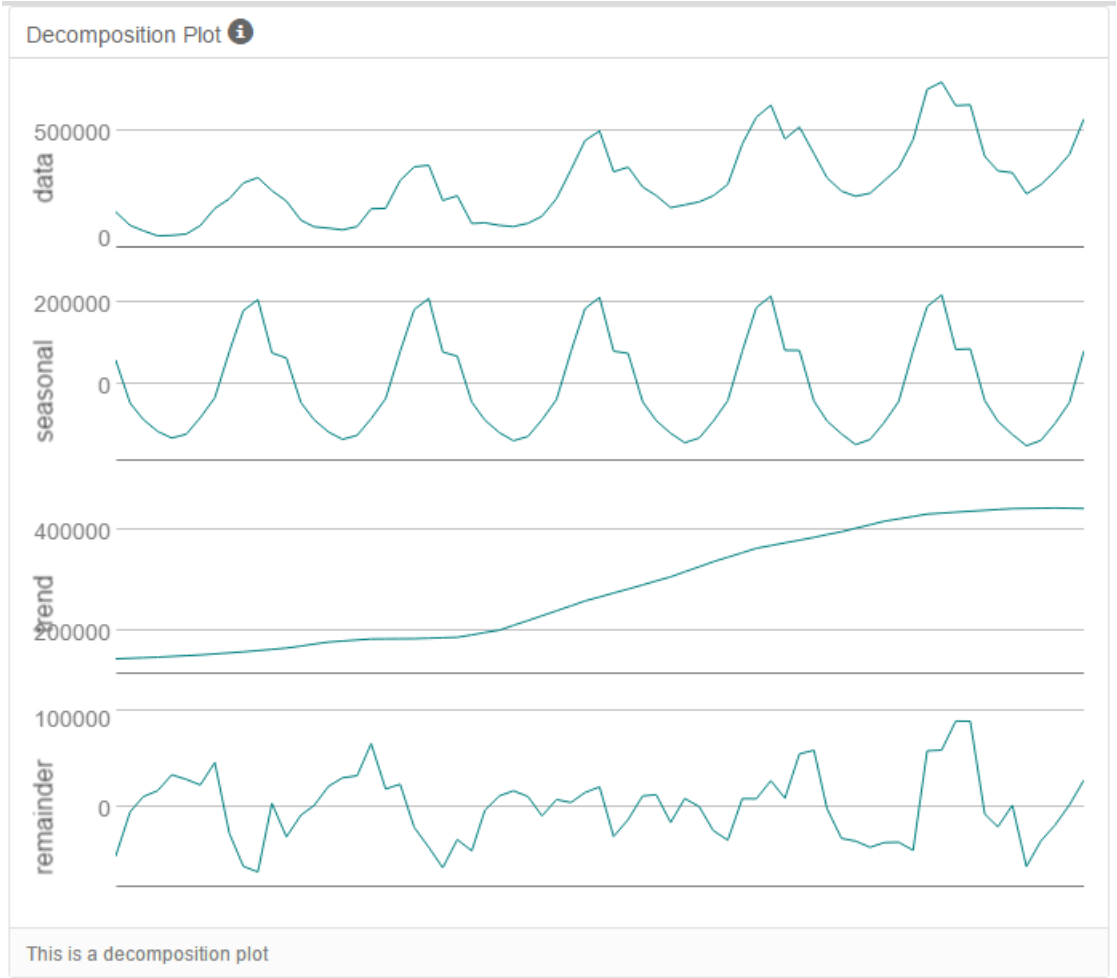
Record #	RecordID	Monthly Sales	Year	Month
1	1	154000	2008	January
2	2	96000	2008	February
3	3	73000	2008	March
4	4	51000	2008	April
5	5	53000	2008	May
6	6	59000	2008	June
7	7	95000	2008	July
8	8	169000	2008	August
9	9	210000	2008	September
10	10	278000	2008	October
11	11	301000	2008	November
12	12	245000	2008	December

Good practice dictates that we should use a holdout (or validation) sample at least as big as the number of periods we are trying to forecast. The business requires a 4-month forecast, so our holdout sample will be records 66 to 69.

Step 2: Determine Trend, Seasonal, and Error components

Using the TS Plot tool in Alteryx we can decompose our time series into its three fundamental components, i.e. trend, seasonality and error (remainder).





The first graph is our time series before being decomposed which means it is a plain graph of our two variables, which are time (year + month) and monthly sales.

Trend: the second graph shows that our monthly sales follow a steadily upward trend except for a ‘plateau’ from July-2009 to Jan-2010 where the increase in sales is minimum.

Seasonality: looking at the third graph, we note that all the peaks are in November and all the valleys are in May. It shows a seasonal pattern repeating every twelve months. Note that peaks and valleys are increasing throughout the years, e.g. Nov2008 is $207467.35 whereas Nov2012 is $219237.38.

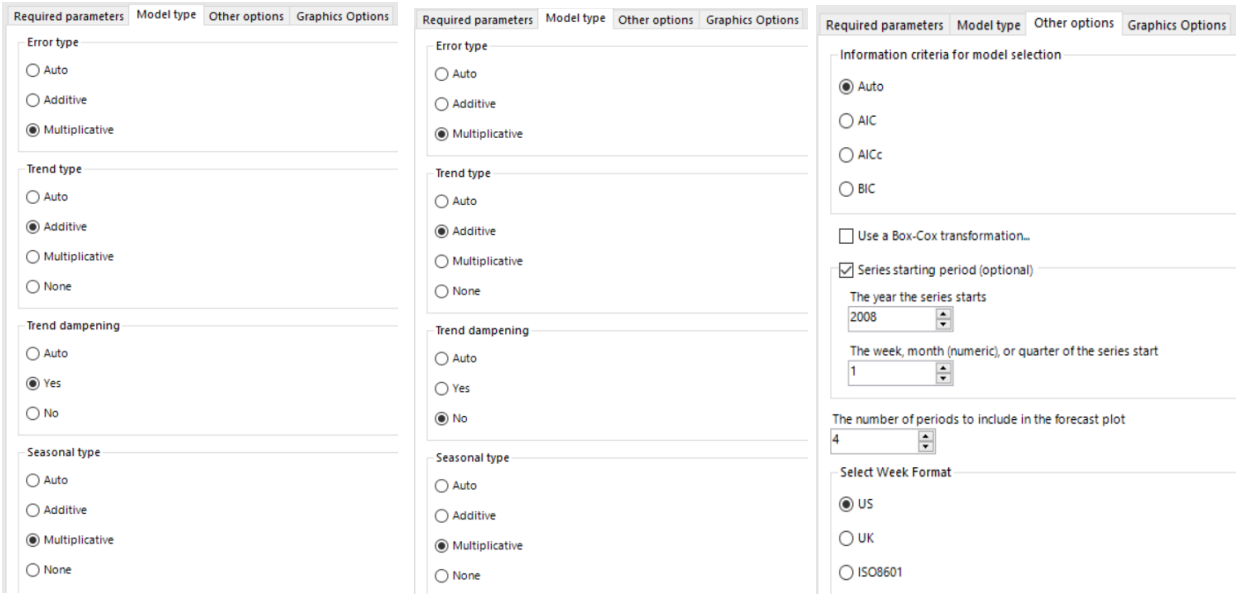
Error: the error is shown in the fourth graph, labelled 'remainder'. The graph does not show a constant variance, as peaks and valleys are bigger towards both ends of the graph than in the middle section.

Step 3: Build your Models

To construct an ETS model we need to determine how to apply the error, trend and seasonality component. Based on the graphs produced for Step 2, it appears that generally the error does not have a constant variance over time. Peaks and valleys are bigger towards both ends of the graph, and smaller in the middle. Hence, we choose to apply the error multiplicatively (M). There is a linear upward trend, so it will be applied additively (A) to our model. Seasonality is the hardest to determine as it appears to be constant over time. However, a closer look at the peaks show a growing magnitude of sales even if it is a slight growth. So, the seasonality component will be applied multiplicatively (M).

There is also the option to include a damper in our ETS (M, A, M) model however we will need to compare the damped and undamped models at the end to see which one yields better predictions.

The series starting period is Jan-2008 and we want to predict 4 periods (also same as holdout).



The damped model has a lower AIC (damped 1639.465 vs undamped) and lower MASE (damped 0.3675478 vs undamped 0.372685), so we are going to choose the **damped ETS(M,A,M)** model over the undamped.

Errors for the damped ETS(M,A,M) model are shown below. We note that MASE is much lower than 1, which generally means it is a 'good' prediction model. We cannot yet comment on the RMSE as we need to compare it to the ARIMA RMSE.

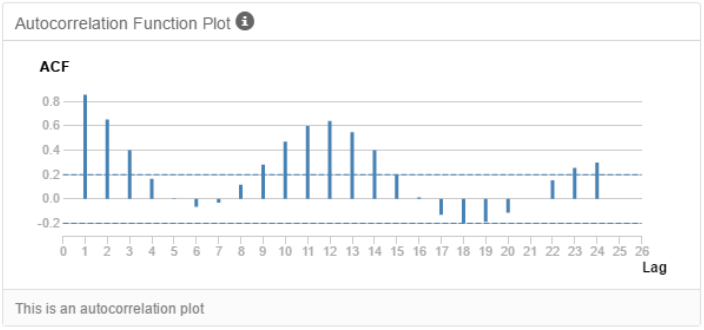
In-sample error measures:

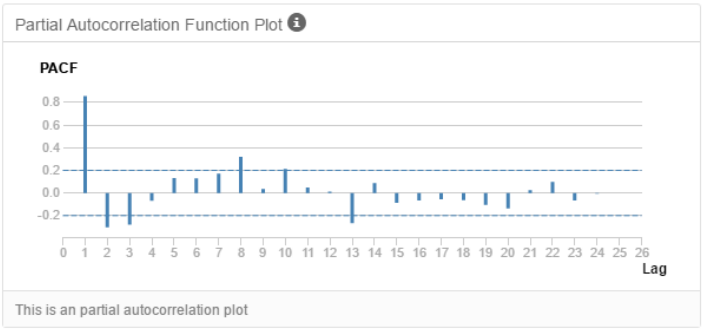
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277

Information criteria:

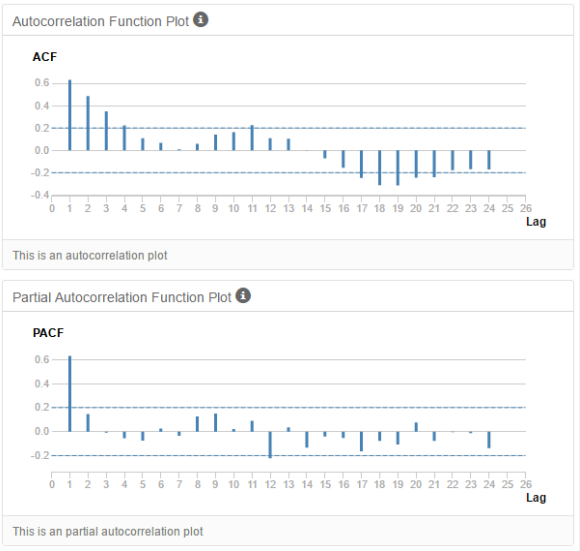
AIC	AICc	BIC
1639.465	1654.3346	1678.604

Since our time series has seasonality, as shown in Step 2, we will need to construct an ARIMA(p,d,q)(P,d,Q)m model. Hence, there are seven model terms to quantify. The seasonal period is 12 months, so m=12. Looking at the ACF and PACF graphs:

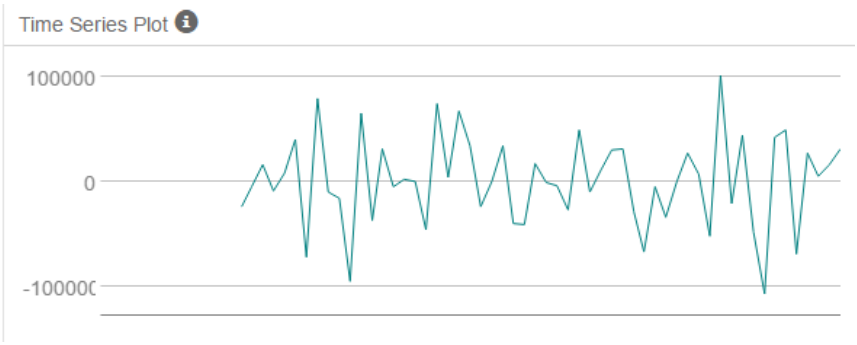




It is obvious that there is serial correlation as there are seasonal increases at the 12 and 24 lags after steadily reducing to 0 in between, so the series needs to be differenced to become stationary. The seasonal difference returns the same 'trend'/pattern, as shown below.

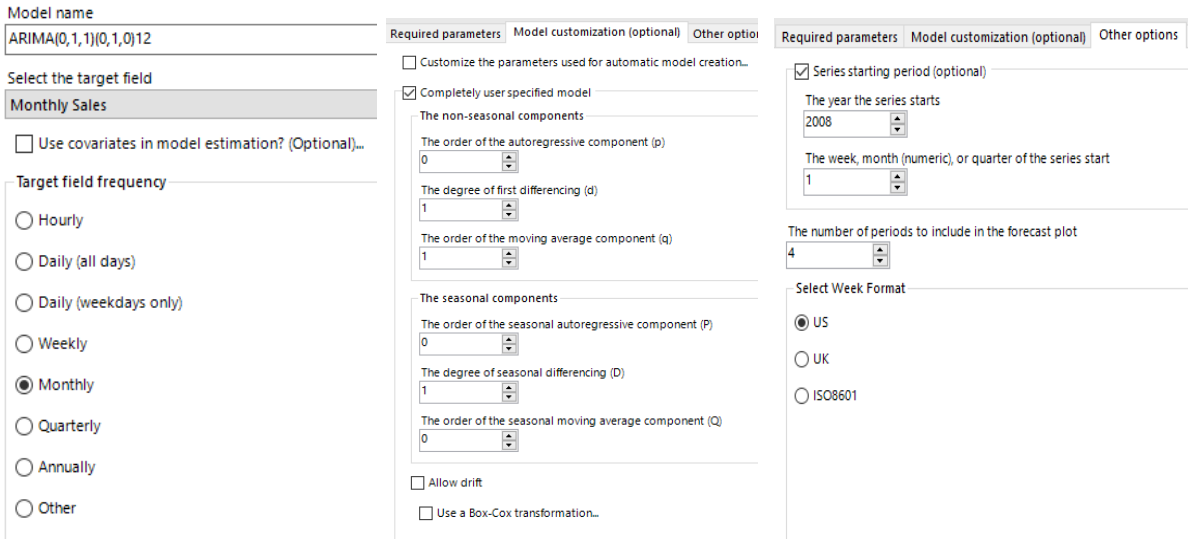


Differencing the seasonal difference, we finally get a stationary time series:



The ACF lag-1 term and is negative and the a sharp cut off is seen so p=0 and q=1, i.e. add an MA term. Seasonal components: the ACF lag-1 is negative so we may not consider adding any SAR tems -> P=0. All seasonal lags (12, 24) do not show a spike so no need to add an SMA term -> Q=0. We used non-seasonal and seasonal differencing, so d=1 and D=1.

We will configure our ARIMA model as an **ARIMA(0,1,1)(0,1,0)12**.



Information Criteria:

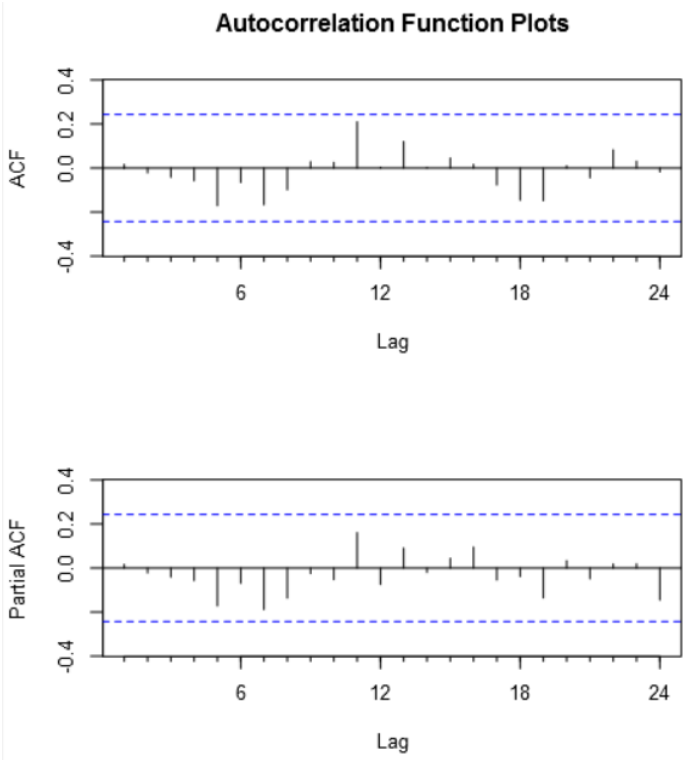
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

ARIMA's RMSE is slightly bigger than ETS's RMSE, but ARIMA's MASE is marginally lower and considerably lower than 1, which indicates that is a 'good' model.

Regraphing of ACF and PACF ::



Step 4: Forecast

As seen in Step 3 the ETS and ARIMA models' in -sample error is:

For the ETS(M,A,M):

Method:
ETS(M,Ad,M)

In-sample error measures:

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

ARIMA(0,1,1)(0,1,0)12

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

Looking at the RMSE, ETS fits better to the time series but ARIMA has a slightly better performance when considering MASE.

Running a forecast for the holdout sample:

1

Comparison of Time Series Models

2

Actual and Forecast Values:

Actual	ETS_Model_MAM_Damped	ARIMA_0_1_1__0_1_0_12
271000	255966.17855	263228.48013
329000	350001.90227	316228.48013
401000	456886.11249	372228.48013
553000	656414.09775	493228.48013

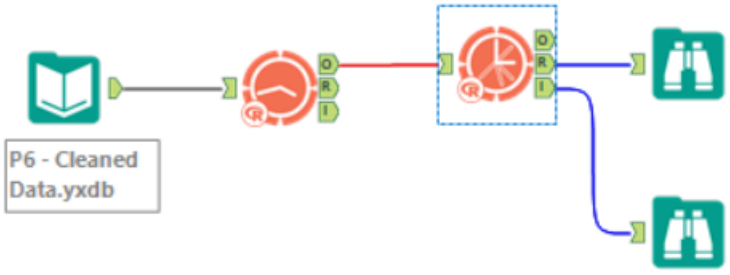
3

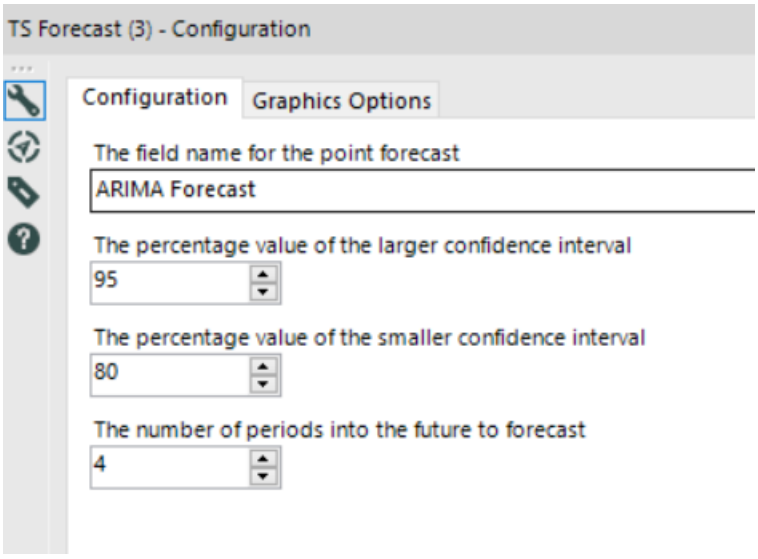
Accuracy Measures:

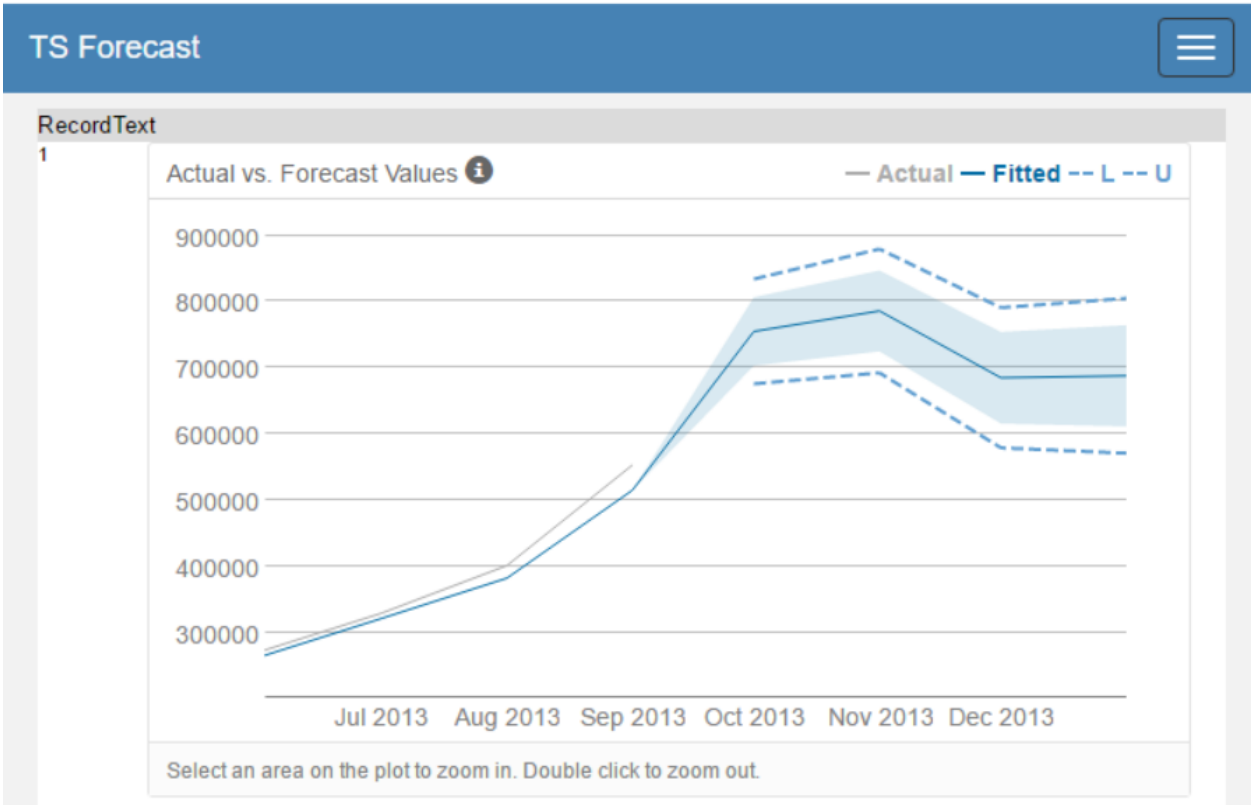
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Model_MAM_Damped	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116	NA
ARIMA_0_1_1__0_1_0_12	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532	NA

The ARIMA model clearly outperforms the ETS model since it has predicted a value closer to the original in every single forecasted period point. Also, the ARIMA model exhibits much lower error measurements especially for RMSE and MASE. Thus, the ARIMA model is the best model to use for forecasting the next four months of videogames demand.

Forecast:







Period	Sub_Period	ARIMA_Forecast	ARIMA_Forecast_high_95	ARIMA_Forecast_high_80	ARIMA_Forecast_low_80	ARIMA_Forecast_low_95
2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462