Project: Forecasting Sales

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.

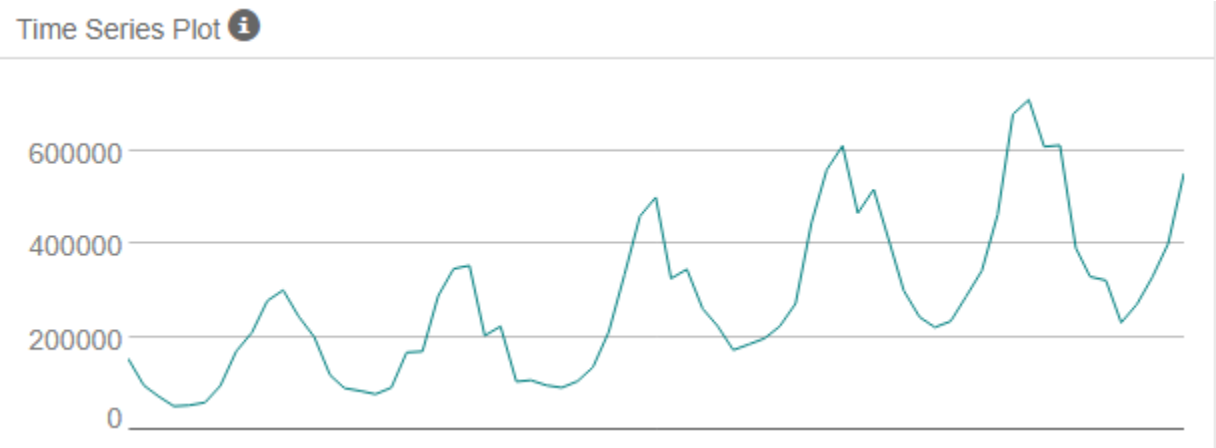

Figure 1. Time series plot for monthly sales of video games in the store.

The time order matters in a time series, and there is a dependency on time so changing the order of values will alter the results.

We can observe the four characteristics:

- the timeseries covers a continuous interval of time
- sequential measurements across the interval of time
- equal spacing between two consecutive measurements
- at least one data point within each time unit of the interval (month).

2. Which records should be used as the holdout sample?

Since the models are going to forecasting four periods, the hold out sample should contain four periods as well, concretely the last four periods (June 2013 to September 2013).

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.

The three main ccomponents — seasonality, trend and error —are shown in Figure 2.

There is a clear seasonality (first plot in Figure 2) December starts with a decreasing trend until May, and then there is an upward trend until November (the peak). The trend line (middle plot in Figure 2) is considered deseasonalized and with an increase in magnitude overtime. Finally, the error or remainder (last plot in Figure 2) is the difference between the value and the trend line estimated.

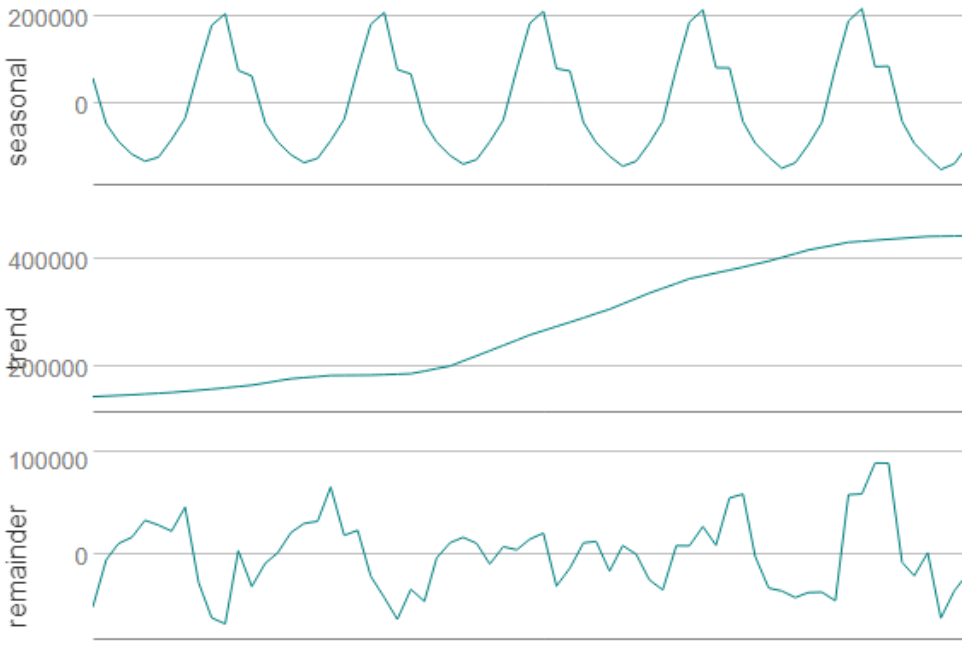

Figure 2. Time series decomposition plot. From top to bottom: seasonal patterns, general tendencies and error.

This time series decomposition plot was obtained with the TS Plot in Alteryx, with the target field Monthly Sales, monthly as frequency and the type of plot called time series plot.

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

On one hand, the error increases variance as the time increases; therefore we should apply the error part multiplicatively. The trend moves linearly so we should apply trend additively. On the other hand, the seasonality increases in volume each seasonal period, so we should apply seasonality in a multiplicative manner. Our model is ETS (M,A,M), an abbreviation of Error Trend Seasonality (Multiplicative, Additive, Multiplicative). Trend dampening is assigned to no. The target field is Monthly Sales, with a monthly frequency. The time series starts in 2008 and the number of periods to include in the forecast plot is 4 (Figure 3).

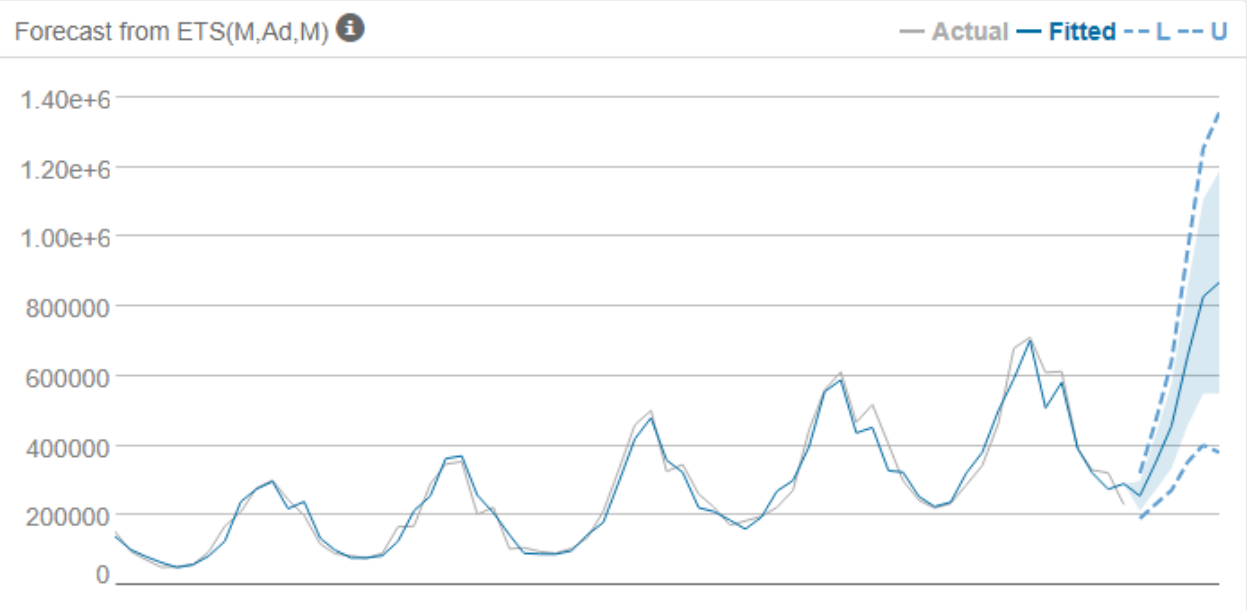

Figure 3. Actual and forecasted values from ETS (M,A,M) model.

The in-sample measures are shown in Figure 4.

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

Figure 4. In-sample measures for the ETS(M,A,M) model.

Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between predicted values and observed values, in the training sample. It shows how many deviations from the mean the forecasted values fall. The value is 33,153.53, which is acceptable, looking at the values in Figure 3.

Mean Absolute Scaled Error (MASE) is defined as the mean absolute error of the model divided by the mean absolute value of the first difference of the series. It measures the relative reduction in error compared to a naive model. A MASE value lower than 1 makes the model acceptable. In our case MASE value is 0.3675478.

The ETS(M,A,M) model with trend dampening has the same in-sample error measures but a lower AIC (Akaike Information Criterion) so it is the preferred method.

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

Figure 5 contains the in-sample measures for the ARIMA model.

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

Figure 5. In-sample measures for the ARIMA(0,1,1)(0,1,0)[12] model.

The RMSE value is 36,761.5281724, which is acceptable, and the MASE is much lower than 1, 0.3646109, which means that ARIMA is an accurate model. The AIC value for the ARIMA model is 1256.5967, lower than the AIC from the ETS model (1639.465).

b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.

Figure 6 shows the autocorrelation function plot (ACF) and partial autocorrelation function plot (PACF) from the original time series. There is a gradual decay with positive autocorrelation in the firsts three lags. PACF cuts off at lag 1 (significant lag).

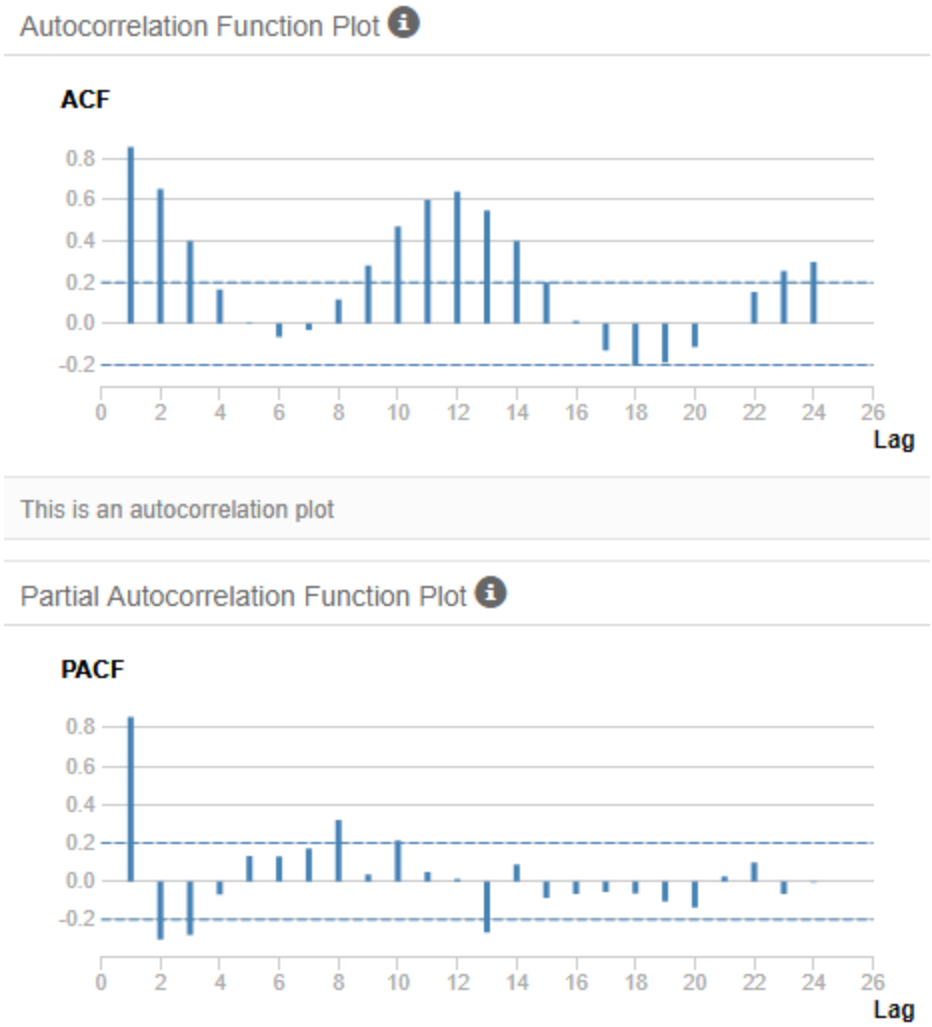

Figure 6. Autocorrelation plots from the original time series.

The seasonal difference time series plot is shown in Figure 7, which is not adjusted for the effect of seasonality. The ACF and PACF plots after seasonal difference (Figure 8) are very similar to the ones of the original time series (Figure 6), with just a slightly reduction in correlation.

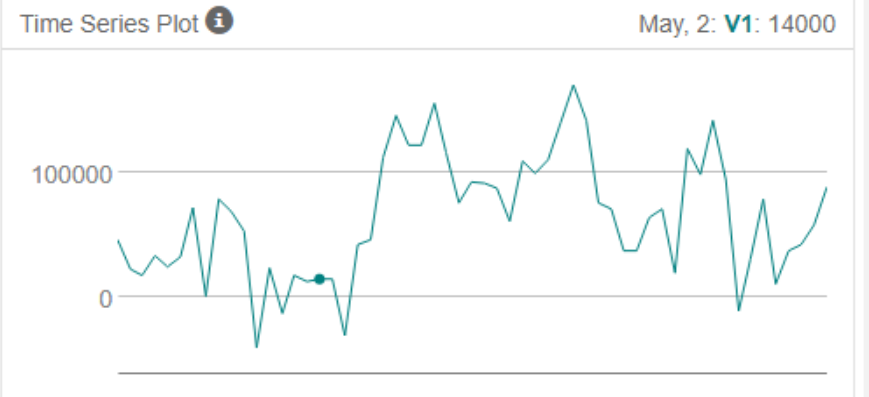

Figure 7. Seasonal difference plot.

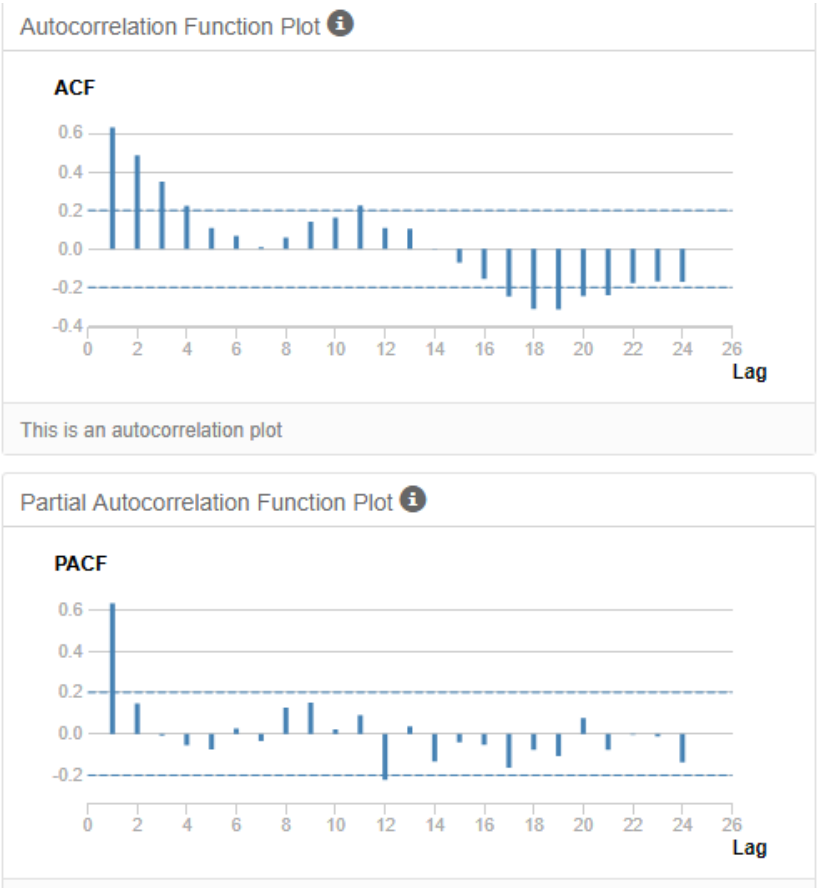

Figure 8. Autocorrelation plots after seasonal difference.

Therefore, to remove correlation we will need to compute the first difference of the seasonal difference. Figures 9 and 10 show that the timeseries and the ACF and PACF, respectively.

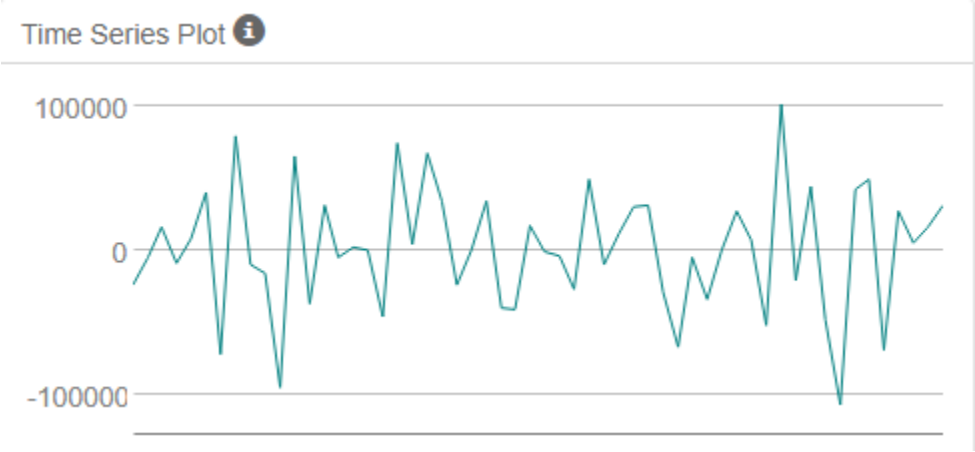

Figure 9. Plot of the first difference of the seasonal difference.

Lag 1 is the only lag left to make the series stationary. Since the autocorrelation is negative (Figure 10) at lag 12 (monthly) we should apply an MA term to make the correction.

Figure 10. Autocorrelation plots after first difference of seasonal difference.

ARIMA needs a MA term, and lower case d of 1 and upper case D of 1 since we took a seasonal difference and the first difference of it: **ARIMA(0,1,1)(0,1,0)[12]**. In Figure 11 we can see that this model makes the time series stationary, so no other terms in the ARIMA model are needed.

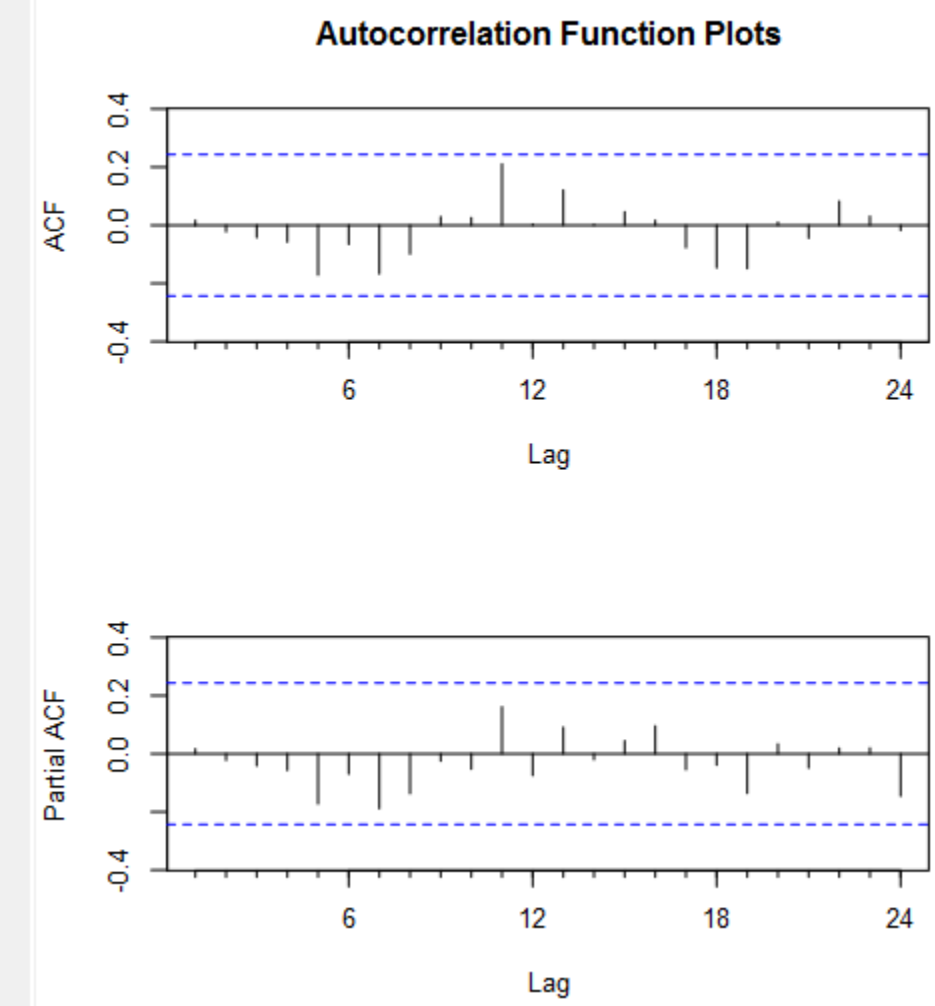

Figure 11. Autocorrelation plots after applying an ARIMA(0,1,1)(0,1,0)[12] model.

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.

Figure 12 contains the in-sample measurements for the ETS model and Figure 13 contains the same for the ARIMA model. MAPE, MASE and ME of the ARIMA model are lower than in the ETS models. This suggests that, on average, the ARIMA model will forecast more accurately. Moreover, AIC was lower in the ARIMA model than in the ETS model (1256.5967 vs. 1639.465).

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

Figure 12. In-sample measures for the ETS(M,A,M) model.

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

Figure 13. In-sample measures for the ARIMA(0,1,1)(0,1,0)[12] model.

Comparing the error measurements for the holdout sample, ARIMA (Figure 15) is performing much better than ETS (Figure 14). ARIMA has a MASE of 0.4532 versus 0.8116 for ETS) and the RMSE value is lower for ARIMA as well (33,999.79 vs. 60,176.47). Therefore, the best model is ARIMA model.

Accuracy Measures:

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

Figure 14. Forecast error measurements for the ETS model.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_video_game	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

Figure 15. Forecast error measurements for the ARIMA model.

Figures 16 and 17 contain the comparison of actual versus forecast values in ETS and ARIMA models, respectively. We can observe that ARIMA values are more close to the actual values.

Actual	ETS_MAM_trend_dampening
271000	255966.17855
329000	350001.90227
401000	456886.11249
553000	656414.09775

Figure 16. Actual and forecast values for the next four periods with the ETS model.

Actual	ARIMA_video_game
271000	263228.48013
329000	316228.48013
401000	372228.48013
553000	493228.48013

Figure 17. Actual and forecast values for the next four periods with the ARIMA model.

2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

Table 1 contains the values, with 95% and 80% confidence intervals and Figure 18 shows the forecast for the next four months (October 2013 to January 2014).

Record #	Period	Sub_Period	forecast_4next	forecast_4next_high_95	forecast_4next_high_80	forecast_4next_low_80	forecast_4next_low_95
1	2013	10	754854.460048	834046.21595	806635.165997	703073.754099	675662.704146
2	2013	11	785854.460048	879377.753117	847006.054462	724702.865635	692331.166979
3	2013	12	684854.460048	790787.828211	754120.566407	615588.35369	578921.091886
4	2014	1	687854.460048	804889.286634	764379.419903	611329.500193	570819.633462

Table 1. Predicted values of monthly sales of video games from October 2013 to January 2014, including 95% and 80% confidence intervals.

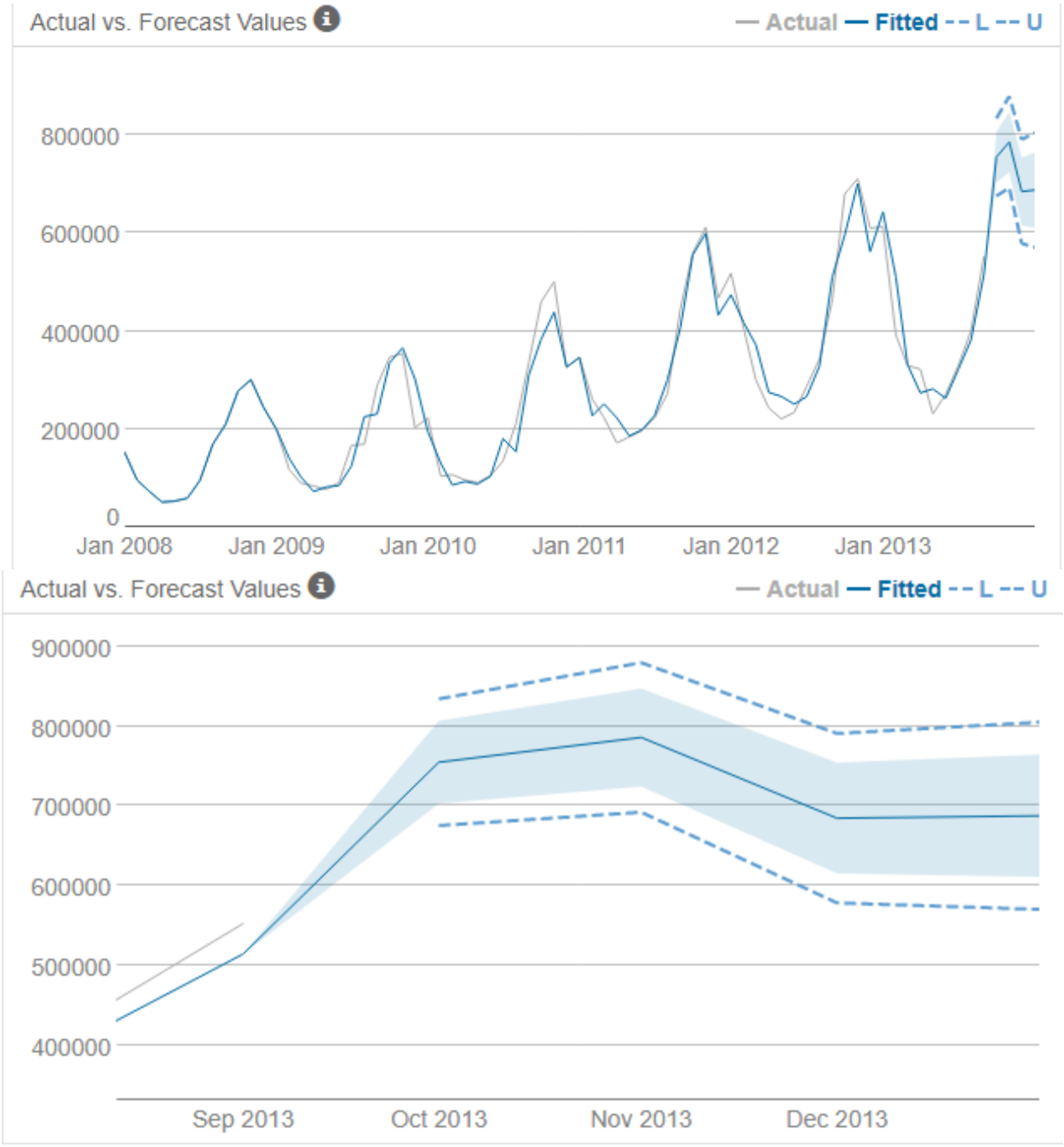

Figure 18. Forecast from October 2013 to January 2014. Top: time series from the actual and fitted data. Bottom: zoomed at the interesting dates.