Kacper Ksieski

# 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.
   Yes, the dataset conforms to each of the four key characteristics of a time series. The dataset is over a continuous time interval from 2008-01 through 2013-09, inclusive. Each measurement takes place in sequence and there is equal spacing of one month between each measurement. Finally, each unit, the month, has at most one data point.
2. Which records should be used as the holdout sample?
   The business has asked for a forecast for the next four months. Therefore, the last four records, which are the most recent periods, should be the holdout sample. These four records take place in a period between 2013-06 and 2013-09, inclusive.

: Awesome: Great, all four key characteristics are explored.

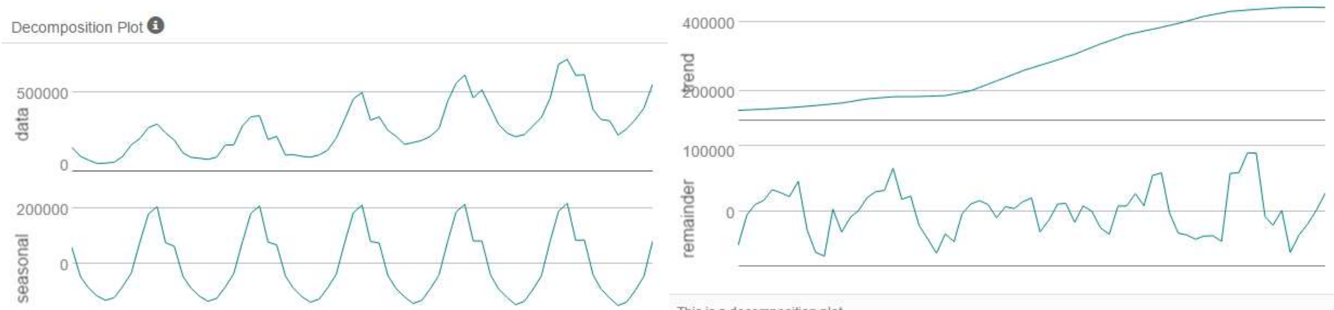
: Awesome; Indeed, we need to use the last 4 records as a holdout sample.

## 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.
   According to the decomposition plot below, the seasonality difference grows in magnitude and is multiplicative. The trend is relatively constant and changes in a linear fashion over time and is additive. The remainder, or error, displays changing variance as the time series moves along and is multiplicative.



This is a decomposition plot

: Awesome: Yes, the seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, even so slightly rather than being constant. In Alteryx, we will need to hover our mouse over the seasonal graph in Interface mode to be able to see that the seasonal numbers are slightly increasing. This is important because:
- Having seasonality suggests that any ARIMA models used for analysis will need seasonal differencing.
- The change in magnitude suggests that any ETS models will use a multiplicative method in the seasonal component.

: Awesome: Correct!

: Awesome: Correct! The error plot of the series presents a fluctuations between large and smaller errors as the time series goes on. Since the fluctuations are not consistent in magnitude then we will apply error in a multiplicative manner for any ETS models

Kacper Ksieski

## 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.
   The model terms for ETS are MAM. The Error term is multiplicative, the Trend term is additive, and the Seasonality term is multiplicative.

   : Awesome: Correct! We have MAM model.

   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
      The in-sample errors are depicted below. The RMSE represents the sample standard deviation of the differences between predicted and observed values. The RMSE of this ETS model is 33153.5267713. The MASE has a value of 0.3675478. MASE errors significantly lower than 1 are ideal. These sample errors will be compared against those of ARIMA to determine the better model.

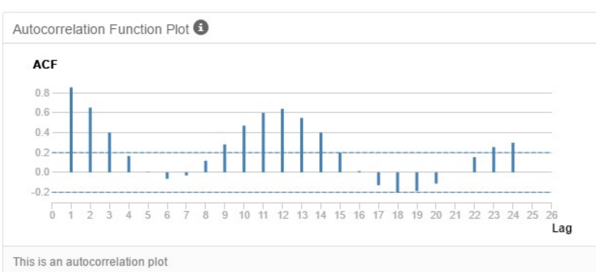
In-sample error measures:

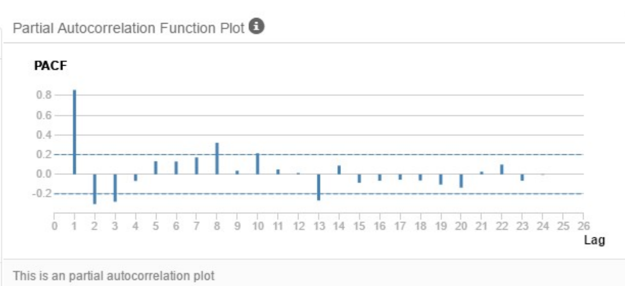
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 5597.130809 | 33153.5267713 | 25194.3638912 | 0.1087234 | 10.3793021 | 0.3675478 | 0.0456277 |

: Awesome: The in-sample errors are correct.

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.
   Depicted below are the ACF and the PACF of the time series. The ACF plot shows seasonality that must be differenced. The ARIMA model will need to be of the form ARIMA(p,d,q)(P,D,Q)[period] because of this seasonality. The period is 12 because the time series is monthly.

   : Awesome: Correct!

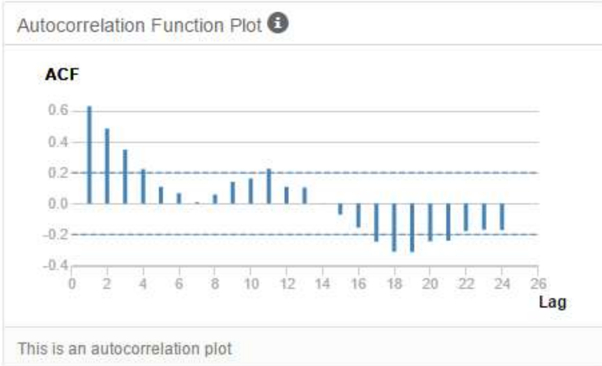


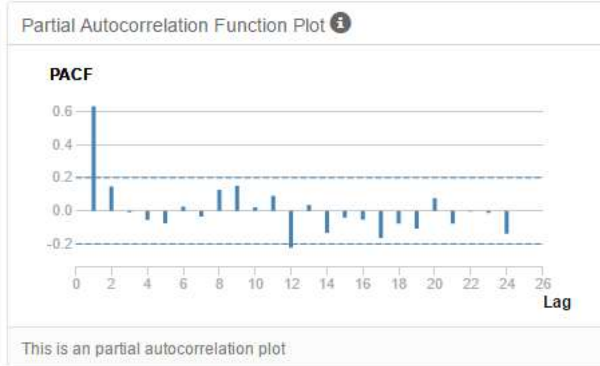




Depicted below are the ACF and the PACF of the time series after taking the first seasonal difference. The time series is not yet stationary. The seasonal difference term is D=1. The ARIMA terms are now ARIMA(p,d,q)(0,1,0)[12].

: Suggestion: More precisely below are depicted Seasonal Difference ACF and PACF plots.
The seasonal difference presents similar ACF and PACF results as the initial plots without differencing, only slightly less correlated. In order to remove correlation we will need to difference further. You are right we have D(1) term.

Kacper Ksieski



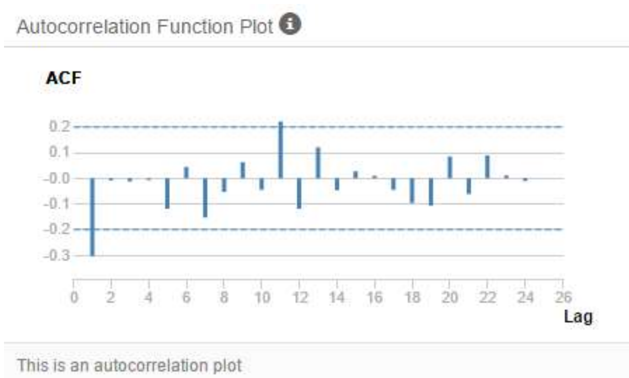



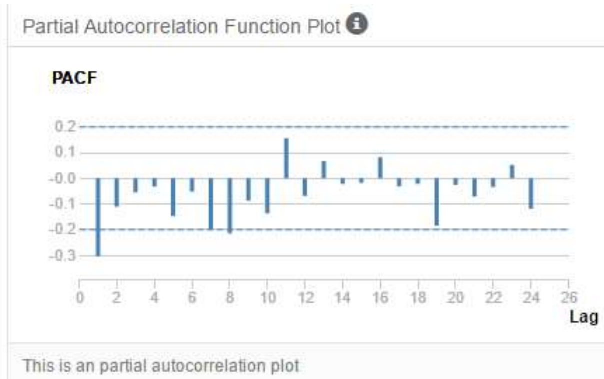

Depicted below are the ACF and the PACF of the time series after the first seasonal difference and the first difference. The plots show that the time series is now stationary. The first difference indicates that d=1. The first bar on the ACF plot is negative, which indicates q=1. Since it is rare to have both a q=1 and a p=1, then p=0. The ARIMA model is now ARIMA(0,1,1)(0,1,0)[12].

: Suggestion: Here again more precisely we have the Seasonal First Difference ACF PACF plots.

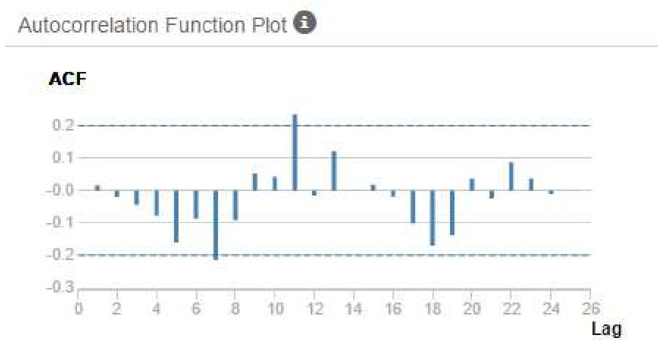
The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

The ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model since there is only 1 significant lag. The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms.

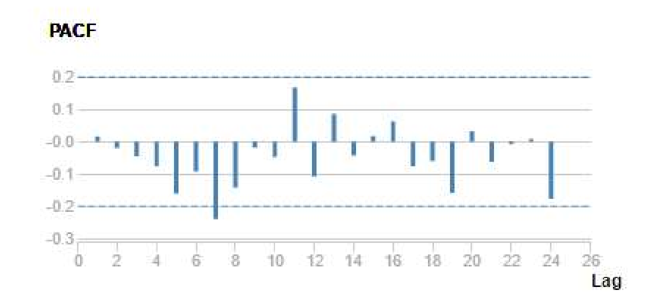



Depicted below are the ACF and the PACF of the time series after the final model terms have been chosen and the entire time series, including the holdout sample, has been used.




: Awesome: Correct the plots here present the final model. The ACF and PACF results for the ARIMA(0, 1, 1)(0, 1, 0)[12] model shows no significantly correlated lags suggesting no need for adding additional AR() or MA() terms.



Kacper Ksieski

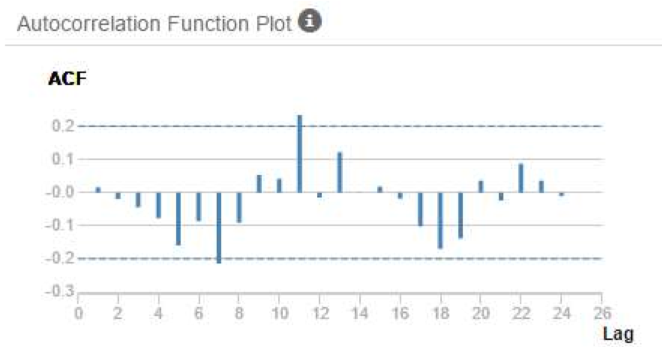
a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
The in-sample errors are depicted below. The RMSE represents the sample standard deviation of the differences between predicted and observed values. The RMSE of this ARIMA model is 36761.5281724. The MASE has a value of 0.3646109. MASE errors significantly lower than 1 are ideal. These sample errors will be compared against those of ETS to determine the better model.

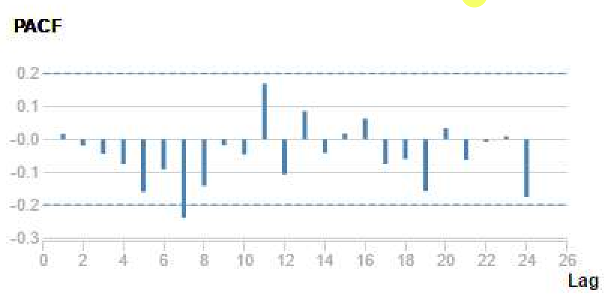
In-sample error measures:

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

: Awesome: The in-sample errors are correct!

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




: Suggestion: These ACF PACF plots are already included above. You can remove them from here.

Kacper Ksieski

## 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.
   I chose the ARIMA(0,1,1)(0,1,0)[12] model.

   Method: ARIMA(0,1,1)(0,1,0)[12]

   Depicted below are the Actual values of the holdout sample compared against the Forecast values of both models.

   Actual and Forecast Values:

| Actual | ETS | ARIMA |
|---|---|---|
| 271000 | 255966.17855 | 263228.48013 |
| 329000 | 350001.90227 | 316228.48013 |
| 401000 | 456886.11249 | 372228.48013 |
| 553000 | 656414.09775 | 493228.48013 |

I created an additional chart, shown below, which shows the absolute and relative differences between the Actual values and the ETS and ARIMA forecasted values. The final field shows the better model as determined by the smaller absolute value of the relative difference of each model. The ARIMA model forecasts more accurately.

| Actual | ETS | ARIMA | ETS_Abs_Diff | ETS_Rel_Diff | ARIMA_Abs_Diff | ARIMA_Rel_Diff | Better_Model |
|---|---|---|---|---|---|---|---|
| 271000 | 255966.17855 | 263228.48013 | 15033.82145 | 0.055475 | 7771.51987 | 0.028677 | ARIMA |
| 329000 | 350001.90227 | 316228.48013 | -21001.90227 | -0.063836 | 12771.51987 | 0.038819 | ARIMA |
| 401000 | 456886.11249 | 372228.48013 | -55886.11249 | -0.139367 | 28771.51987 | 0.071749 | ARIMA |
| 553000 | 656414.09775 | 493228.48013 | -103414.09775 | -0.187006 | 59771.51987 | 0.108086 | ARIMA |

Below are the accuracy measures of both models. The ARIMA model beats the ETS model in each measure by having smaller absolute values.

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| ETS | -41317.07 | 60176.47 | 48833.98 | -8.3683 | 11.1421 | 0.8116 | NA |
| ARIMA | 27271.52 | 33999.79 | 27271.52 | 6.1833 | 6.1833 | 0.4532 | NA |

I included the AIC measures for each model as well. The better model is usually the one with the lower AIC score. In this case, the ARIMA model has the better score and this fits with the rest of the information which points to the ARIMA model.

| ETS | ARIMA |
|---|---|
| Akaike Info. Criterion | Akaike Info. Criterion |
| 1673.4 | 1350 |

: Awesome: Correct! The final better model is ARIMA(0, 1, 1)(0, 1, 0)[12].

: Awesome: Great comparison here! Yes, we can see that ARIMA is the better model looking at the forecasts as well.

: Awesome: Great job showing the accuracy measures. When looking at the model's ability to predict the holdout sample, we see that the ARIMA model has better predictive qualities in just about every metric.

: Awesome: Yes, ARIMA has a lower AIC.

: Required: Excellent comparison here. Just please note that when choosing the better model we also need to take into account the in-sample errors. I am referring to the same in-sample errors shown in Step 3. We need to show and compare the in-sample errors as well.

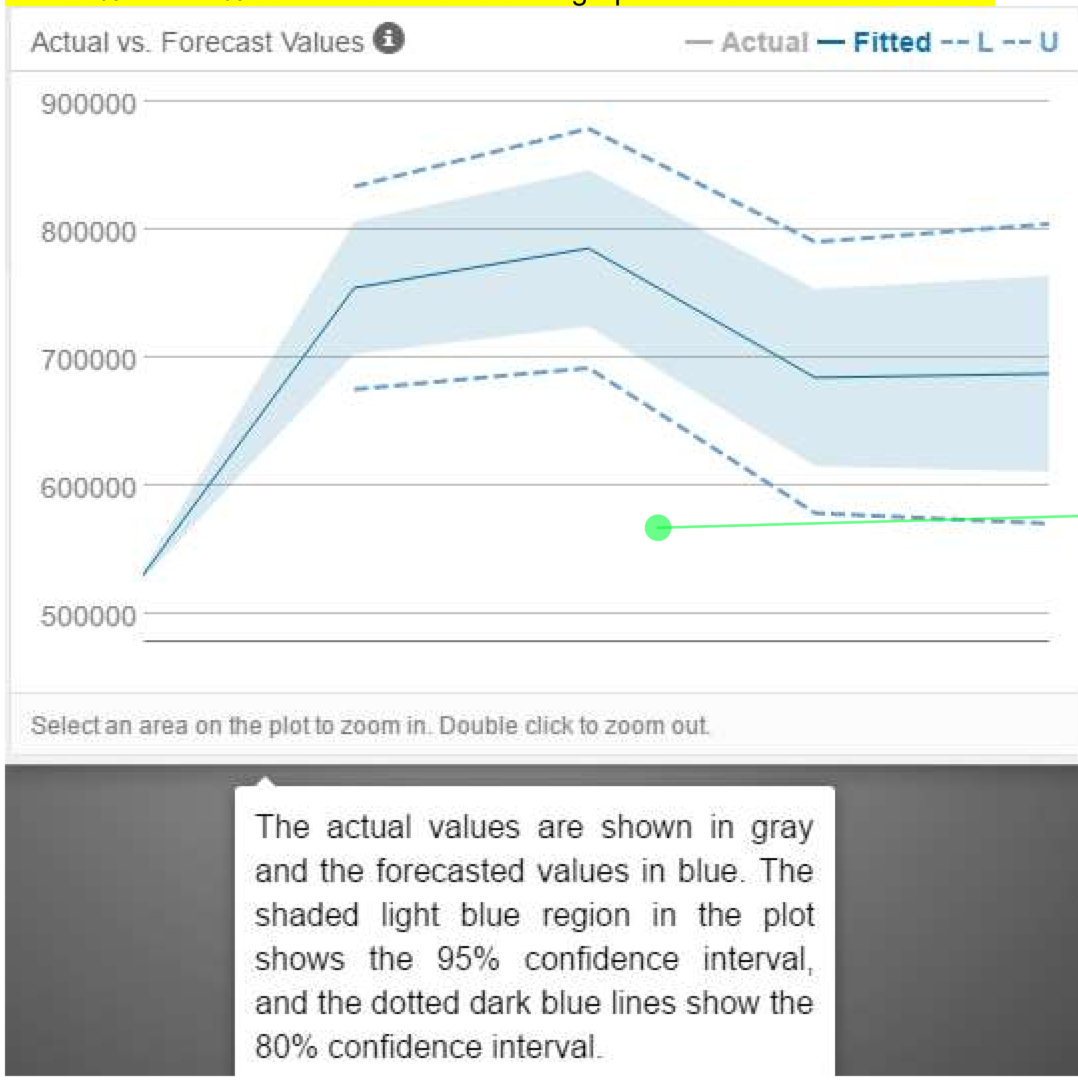
Kacper Ksieski

2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.
The forecast for the next four periods is shown below, in addition to the 80% and 95% confidence intervals.

| Period | Sub_Period | forecast | forecast_high_95 | forecast_high_80 | forecast_low_80 | forecast_low_95 |
|---|---|---|---|---|---|---|
| 6 | 10 | 754854.460048 | 834046.21595 | 806635.165997 | 703073.754099 | 675662.704146 |
| 6 | 11 | 785854.460048 | 879377.753117 | 847006.054462 | 724702.865635 | 692331.166979 |
| 6 | 12 | 684854.460048 | 790787.828211 | 754120.566407 | 615588.35369 | 578921.091886 |
| 7 | 1 | 687854.460048 | 804889.286634 | 764379.419903 | 611329.500193 | 570819.633462 |

: Awesome: Perfect! The forecasts are correct!

The 80% and 95% confidence intervals are graphed below with the forecast.




: Awesome: Excellent work with the plot!

## Before you Submit

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