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:

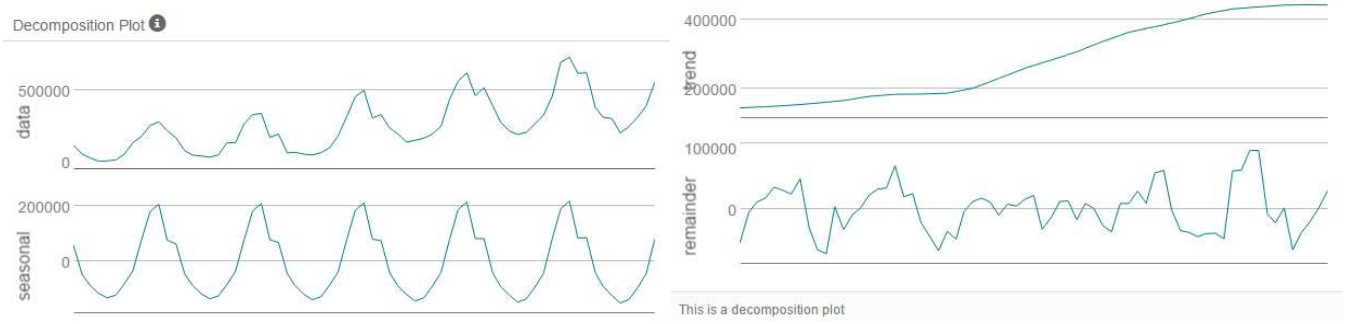
1. Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.
   Yes, the dataset conforms to each of the four key characteristics of a time series. The dataset is over a continuous time interval from 2008-01 through 2013-09, 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.

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.



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)

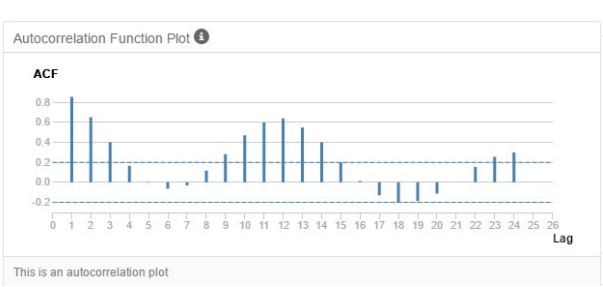
Answer these questions:

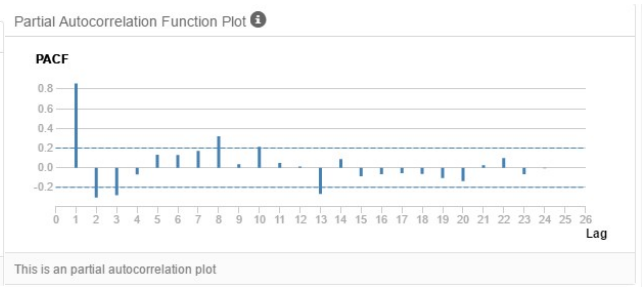
1. What are the model terms for ETS? Explain why you chose those terms.
   The model terms for ETS are MAM. The Error term is multiplicative, the Trend term is additive, and the Seasonality term is multiplicative.
   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
      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

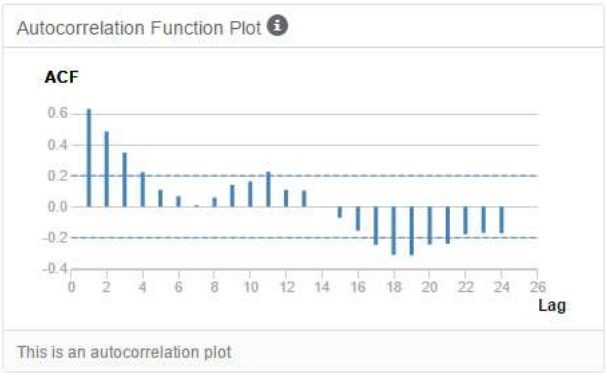
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
   Depicted below are the ACF and 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.

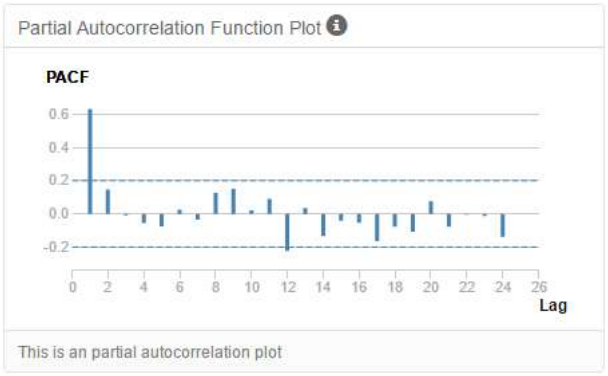




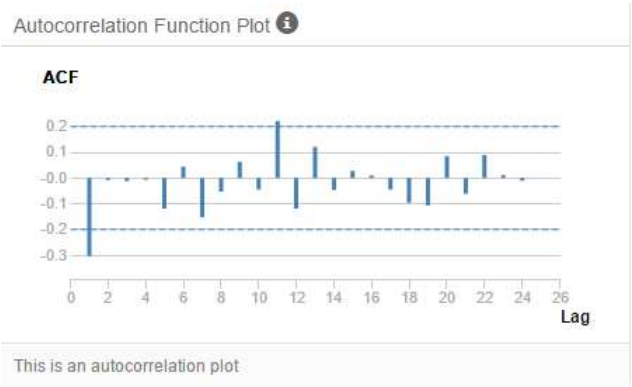
   Depicted below are the Seasonal Difference ACF and the PACF plots. The time series still exhibits correlation and is not yet stationary. The seasonal difference term is D(1). The ARIMA terms are now ARIMA(p,d,q)(0,1,0)[12].

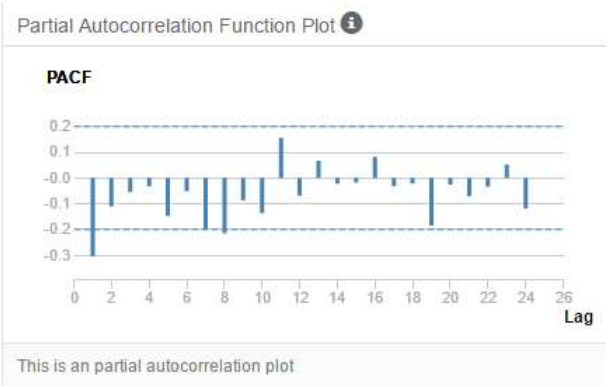
Kacper Ksieski



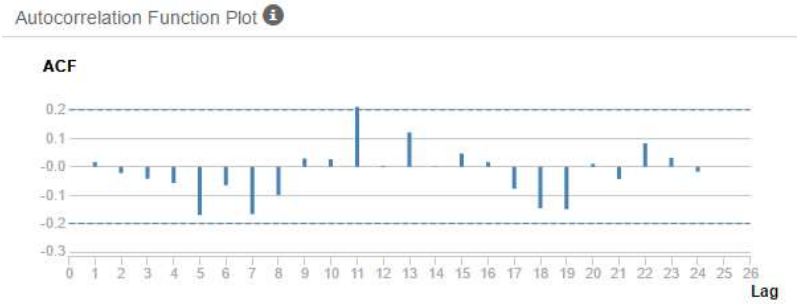


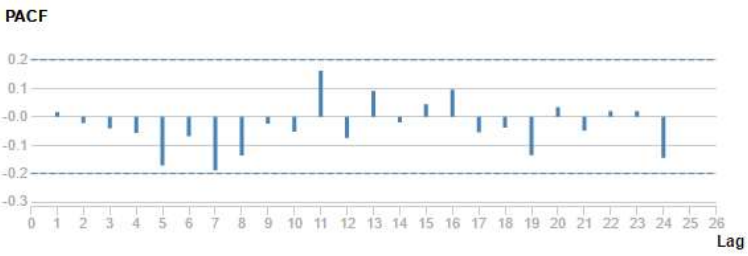
Depicted below are the Seasonal First Difference ACF and the PACF plots. The plots show that the time series is now stationary. The first difference indicates a d(1) term and there is no need for further differencing. Lag 1 on the ACF plot is negative, which indicates an MA model and a q(1) term. Since it is rare to have both a q(1) term and a p(1), then p(0). This is confirmed by the lack of significant correlation at the seasonal lags (12, 24, etc.). The ARIMA model is now ARIMA(0,1,1)(0,1,0)[12].





Depicted below are the ACF and the PACF of the time series after the final model terms have been chosen. The holdout sample has not been included.





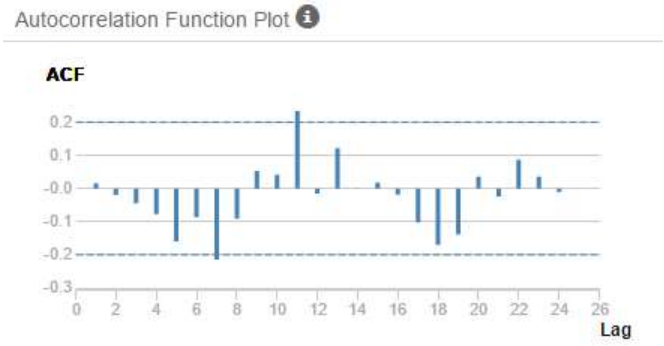
Kacper Ksieski

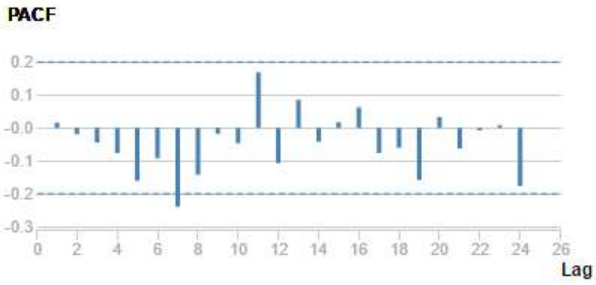
a. Describe the in-sample errors. Use at least RMSE and MASE when examining results
The in-sample errors are depicted below. The RMSE represents the sample standard deviation of the differences between predicted and observed values. The RMSE of this ARIMA model is 36761.5281724. The MASE has a value 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

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 plots for the ARIMA model for the time series and include all data points including the holdout sample.





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

   Below is a chart depicting the in-sample error measurements. The ARIMA model has smaller absolute values in most of the metrics. The ARIMA model appears to produce smaller predictive errors in more metrics than the ETS model.

MODEL	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
ETS	5597.130809	33153.5267713	25194.3638912	0.1087234	10.3793021	0.3675478	0.0456277
ARIMA	-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145
betterModel	ARIMA	ETS	ARIMA	ETS	ARIMA	ARIMA	ARIMA

   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. This shows the ARIMA model has better predictive powers in just about every metric.

   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

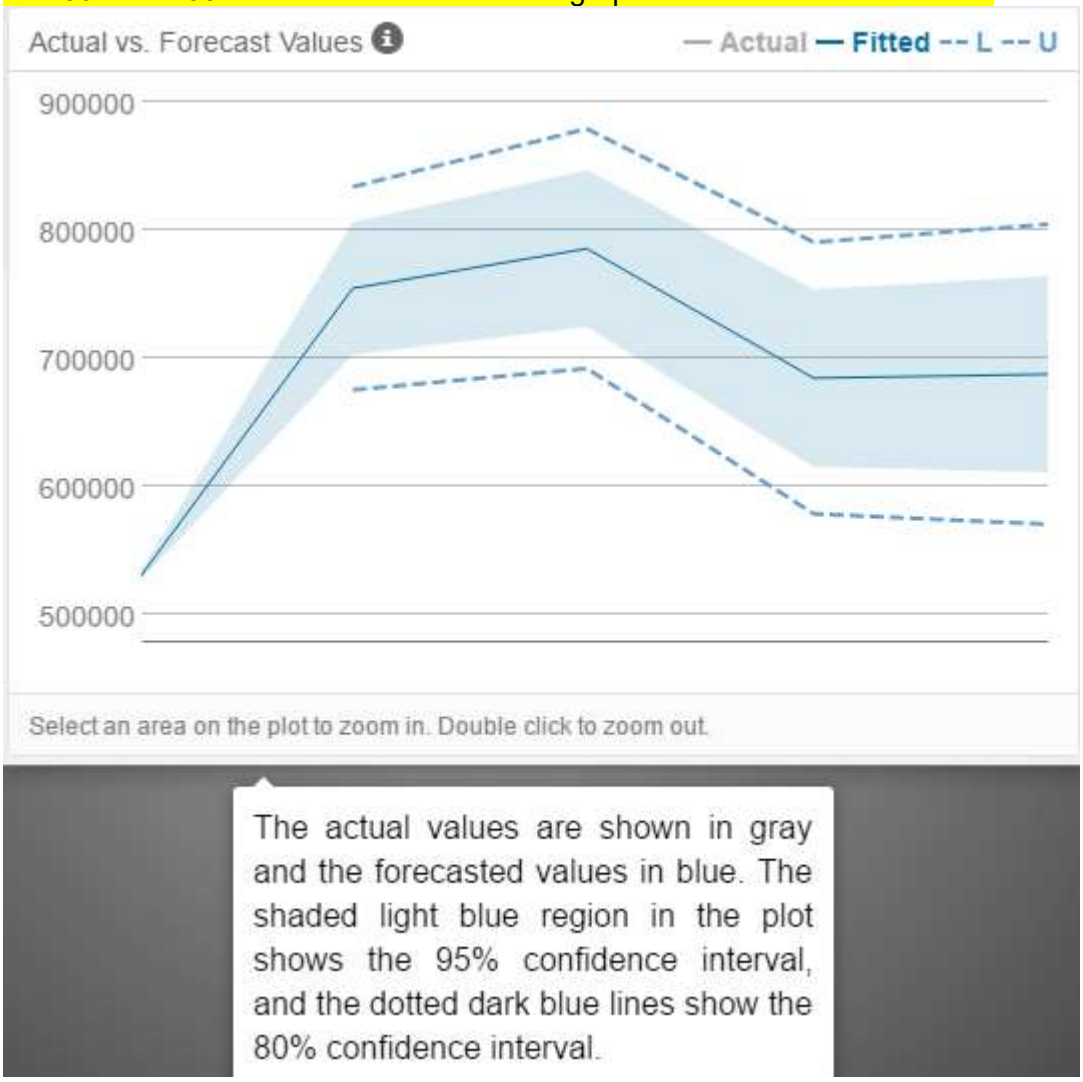
I included the AIC measures for each model as well. The better model is usually the one with the lower AIC score. In this case, the ARIMA model has the better score and this fits with the rest of the information which points to the ARIMA model.

ETS	ARIMA
Akaike Info. Criterion	Akaike Info. Criterion
1639.5	1350

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

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



Kacper Ksieski

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.