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.

   The dataset meets the criteria of a time series dataset. It's over a continuous time interval, the file contains store information for the company's sales by month from 2008-01 to 2013-09; There are sequential measurements across that interval; There is equal spacing between every two consecutive measurements; Each time unit within the time interval has at most one data point (Monthly sales)

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

   I have used the last 4 records to be my holdout sample (2013-06 to 2013-09)

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 trend is uptrend with a regularly occurring spike in sales each year. Seasonality shows that the regularly occurring spike in sales each year changes in magnitude, ever so slightly. The error plot of the series presents a fluctuation between errors as the time series goes on.

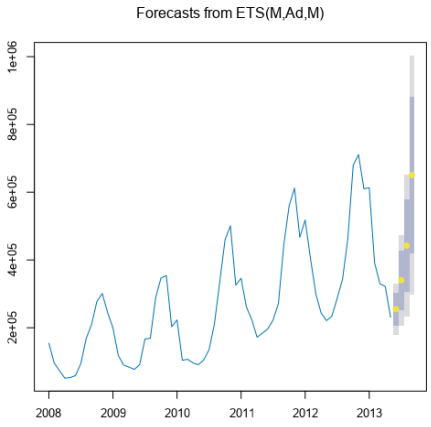
Step 3: Build your Models

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 trend line exhibits linear behavior so I will use an additive method. The seasonality changes in magnitude each year so a multiplicative method is necessary. The error changes in magnitude as the series goes along so a multiplicative method will be used.



The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,Ad,M)

In-sample error measures:

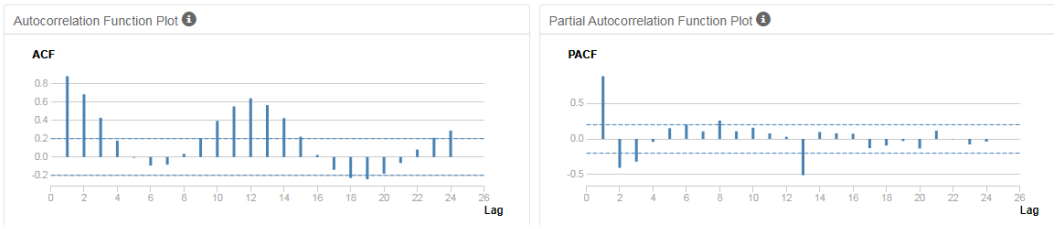
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3243.4703524	31474.3668886	24188.2167878	-0.572395	10.3052041	0.3528697	0.0087402

Information criteria:

AIC	AICc	BIC
1640.1232	1654.9928	1679.2622

RMSE (how close the observed data points are to the model's predicted values) = 31,474.37 units
MASE (is a measure of the accuracy of forecasts) = 0.35 (When MASE = 0.5, means that our model has doubled the prediction accuracy. The lower, the better.)

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.
   Since there are seasonal components found in the time series I will use an ARIMA(p, d, q)(P, D, Q)S model for forecasting.



Seasonal First Difference:
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.

   a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

Summary of ARIMA Model ARIMA

Method: ARIMA(0,0,2)(1,0,1)[12]

Call:
auto.arima(Monthly.Sales, d = 0, D = 0, max.p = 2, max.q = 2, max.P = 1, max.Q = 1, ic = "aicc", allowdrift = TRUE)

Coefficients:

	ma1	ma2	sar1	sma1	intercept
Value	0.756857	0.454069	0.79139	0.333184	279276.830116
Std Err	0.152871	0.093872	0.083079	0.187651	52323.055018

sigma^2 estimated as 2422782992.19822: log likelihood = -801.69185

Information Criteria:

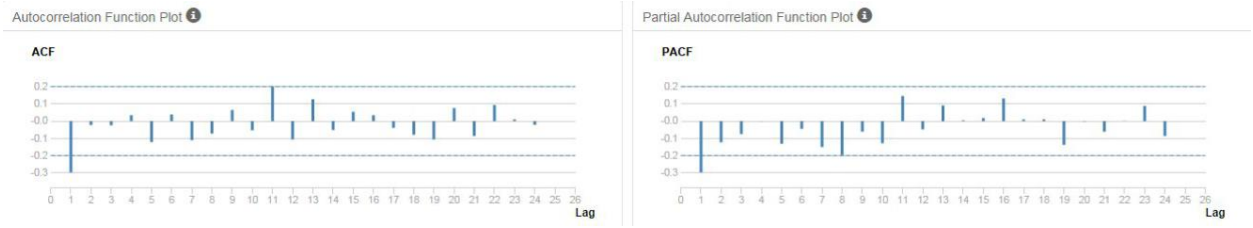
AIC	AICc	BIC
1615.3837	1616.832	1628.43

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
11605.5609017	47290.7503612	37420.2419184	-4.0859432	19.2169236	0.545905	0.1140751

RMSE = 47,290.75
MASE = 0.55

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



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 ETS model because it showed lower values in RMSE and MASE.
   The Information criteria has lower values in the ARIMA model, but hence they are not different by a lot from the ones in the ETS, I choose the ETS model.

Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,Ad,M)

In-sample error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	3243.4703524	31474.3668886	24188.2167878	-0.572395	10.3052041	0.3528697	0.0087402

Information criteria:

AIC	AICc	BIC
1640.1232	1654.9928	1679.2622

Summary of ARIMA Model ARIMA

Method: ARIMA(0,0,2)(1,0,1)[12]

Call:
auto.arima(Monthly.Sales, d = 0, D = 0, max.p = 2, max.q = 2, max.P = 1, max.Q = 1, ic = "aicc", allowdrift = TRUE)

Coefficients:

	ma1	ma2	sar1	sma1	intercept
Value	0.756857	0.454069	0.79139	0.333184	279276.830116
Std Err	0.152871	0.093872	0.083079	0.187651	52323.055018

sigma^2 estimated as 2422782992.19822: log likelihood = -801.69185

Information Criteria:

AIC	AICc	BIC
1615.3837	1616.832	1628.43

In-sample error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
	11605.5609017	47290.7503612	37420.2419184	-4.0859432	19.2169236	0.545905	0.1140751

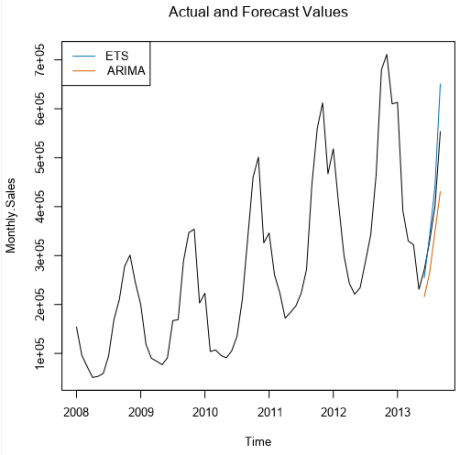
Comparison of Time Series Models

Actual and Forecast Values:

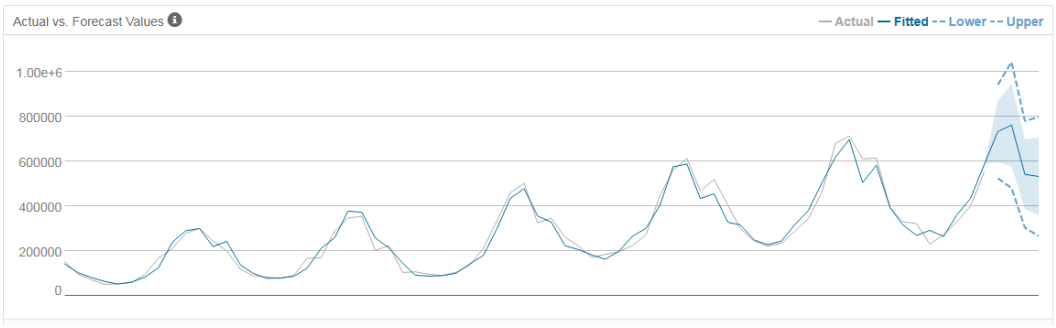
Actual	ETS	ARIMA
271000	254853.70905	216025.73602
329000	340280.41766	264501.38274
401000	442291.20116	348529.623
553000	650453.11029	430799.30678

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-33469.61	53828.48	41542.75	-6.3476	9.3266	0.6904
ARIMA	73535.99	78848.58	73535.99	18.7682	18.7682	1.2221



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



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
6	10	731873.553663	940101.715508	868026.630407	595720.476919	523645.391818
6	11	760721.120998	1041825.301239	944525.264906	576916.97709	479616.940757
6	12	541120.282806	778583.165285	696388.915903	385851.649708	303657.400326
7	1	531689.196619	797032.300895	705187.726885	358190.666353	266346.092343

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.