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.

To meet the criteria of a time series dataset,
1) Each measurement of data should be taken across a continuous time interval
2) The time interval is sequential
3) Each time unit should have at most one data point
4) There should be equal intervals across 2 consecutive data time points.

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

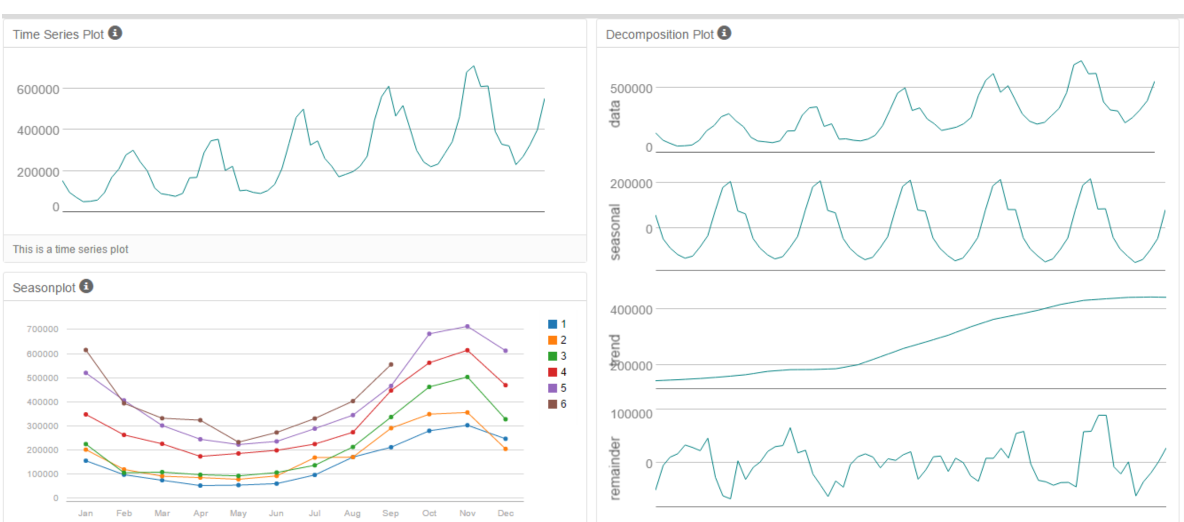
   The first 65 records will be used to construct our forecast model and the last 4 records or the last 4 months of data will be used as the holdout sample.
   in this case from 2015-July to 2015-Dec

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 seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, ever so slightly.
The error plot of the series presents fluctuations between large and smaller errors as the time series goes on.
The time series plot shows an upward rising trend.

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

The chosen model terms for the ETS model are –
For trend – its additive, as the trend line is linear.
For error – its multiplicative as it is not constant and changes continuously.
For seasonality – Looks to be constant but the peaks have slight changes showing an increasing variation in the graph – so also multiplicative.
ETS(M,A,M).

Method:
ETS(M,A,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Information criteria:

AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

Smoothing parameters:

Parameter	Value
alpha	0.787787
beta	1e-04
gamma	0.000522

RMSE=32992.72 and MASE=0.37 and AIC=1639.73.

The in-sample errors we use here are the RMSE (Root mean Square error) – that is the differences in standard deviation of the sample and MASE (Mean Absolute Scaled Error) – is a measure of the accuracy for the forecasts and is recommended for determining comparative accuracy of forecasts.

The lower the value for our RMSE , the better the variances can be explained and for MASE to be considered as 'good' the value should be lower than 1.
In our case the MASE calculated are all below the value 1 but here, we choose the model that has a lower MASE, as the lower MASE means lower calculated error.

Coincidentally,
We do not use the dampened trend as it has a Higher RMSE than its undampened counterpart.

Summary of Time Series Exponential Smoothing Model ETS_project_dampened

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

Information criteria:

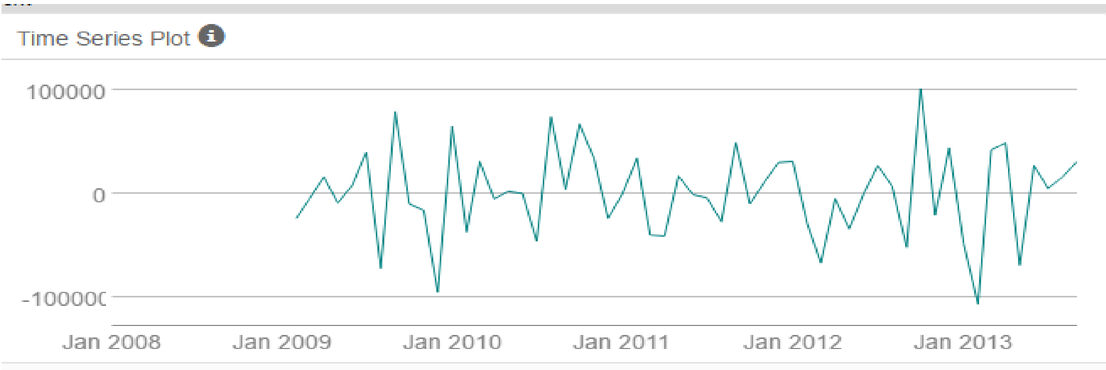
AIC	AICc	BIC
1639.465	1654.3346	1678.604

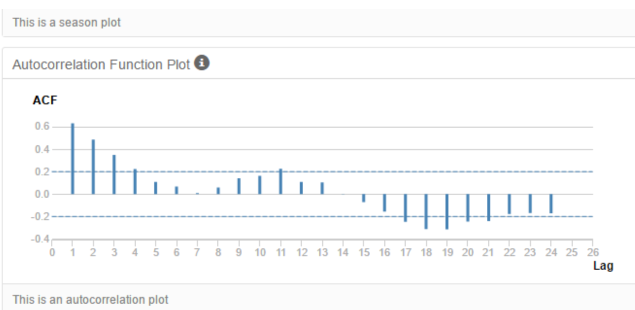
2. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.

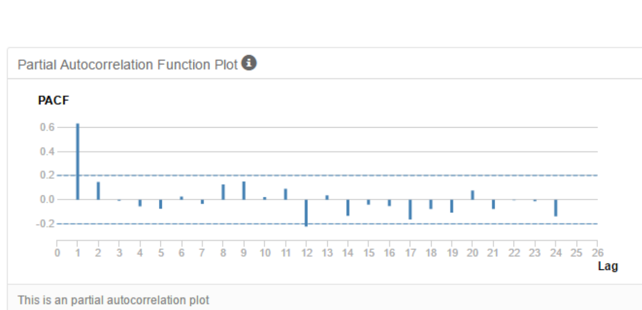
The model terms are non-seasonal (p,d,q) also known as p-auto regressive , d-difference and q is for moving average. For seasonal (P,D,Q)m is the same as the non-seasonal but it's the auto regressive , difference and moving average for the seasonal part.

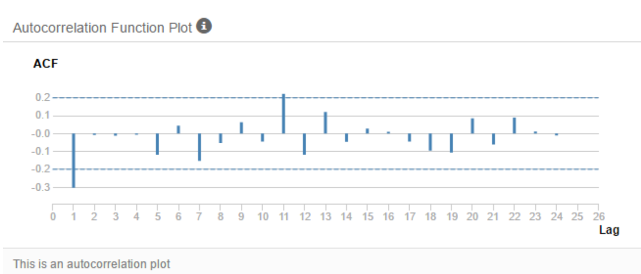
We can see that the Time series plot for the row difference is still not stationary, so we will have to use the 1st seasonal difference to get the stationary time series plot.

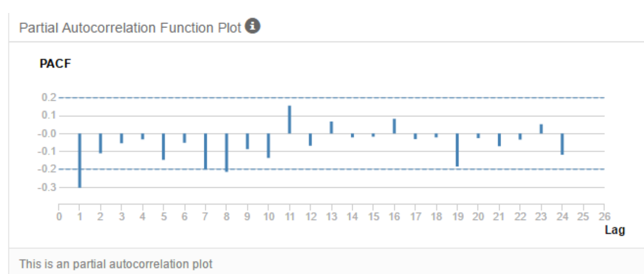
The respective ACF and PACF graphs for the row difference is shown below-











The above two graphs are the result of a 1st seasonal difference and we see that the time series plot is stationary around the value 0.
And observing the respective ACF and PACF plots we can say that the correlation for the lag periods have reduced.

But we can still see that there are high correlations in lag 1 in both ACF and PACF and lag at periods 7 and 8 in PACF graph.

So, we'll have to choose the model components depending on the above ACF and PACF graphs.
Our auto regressive part is 0 in this case so, p=0;
Since we use the first seasonal difference, our d=1;
Since we do not have a negative seasonal lag there is no need to include a seasonal MA term- our moving average is set to 0.
Using this data, we get the table(in the next page) as a result.

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

Report

Summary of ARIMA Model Arima_project_new

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

Call:
Arima(Monthly.Sales, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

In-sample error measures:

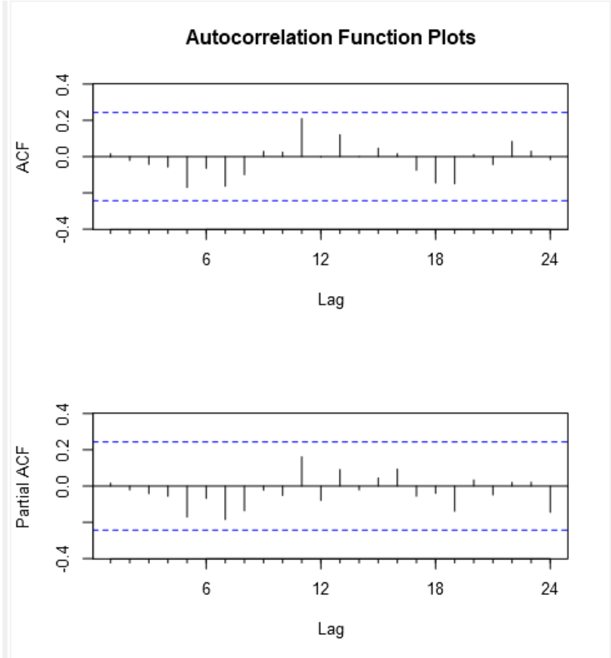
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

Ljung-Box test of the model residuals:
Chi-squared = 16.4458, df = 23, p-value = 0.83553

Our RMSE=36761.5 and MASE=0.36 and AIC=1256.59.

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

Plots



We can see the correlations for the lags have reduced for this ARIMA model and now are stationary.

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.

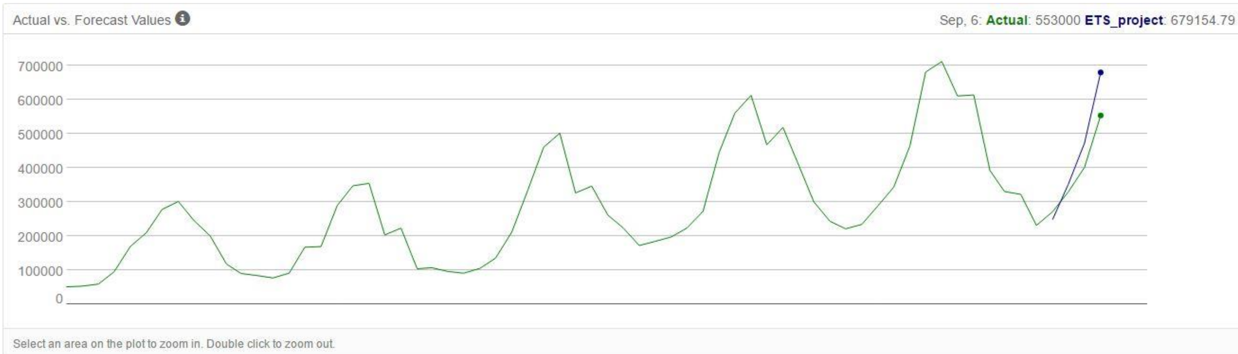
For our ARIMA model -

Our RMSE=36761.5 and MASE=0.36 and AIC=1256.59.

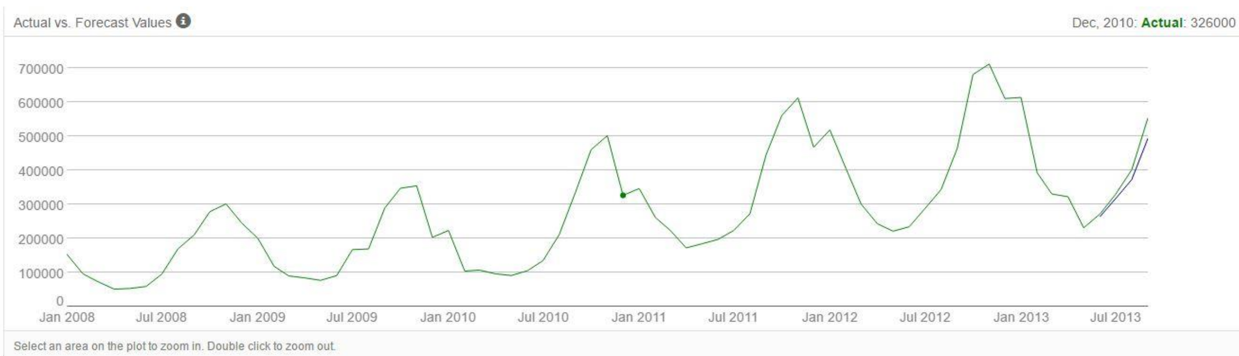
And for our ETS model-

RMSE=32992.72 and MASE=0.37 and AIC=1639.73.

Both the models have similar MASE values and our ARIMA model has a higher RMSE value.



ETS model for the forecast against the holdout sample.



ARIMA model for the forecast against the holdout sample.

From both the above charts we can see that the ARIMA model performed better as the forecasted value for the holdout sample is the closest than the ETS model.

Method:
ETS(M,A,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2818.2731122	32992.7261011	25546.503798	-0.3778444	10.9094683	0.372685	0.0661496

Information criteria:

AIC	AICc	BIC
1639.7367	1652.7579	1676.7012

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

Information Criteria:

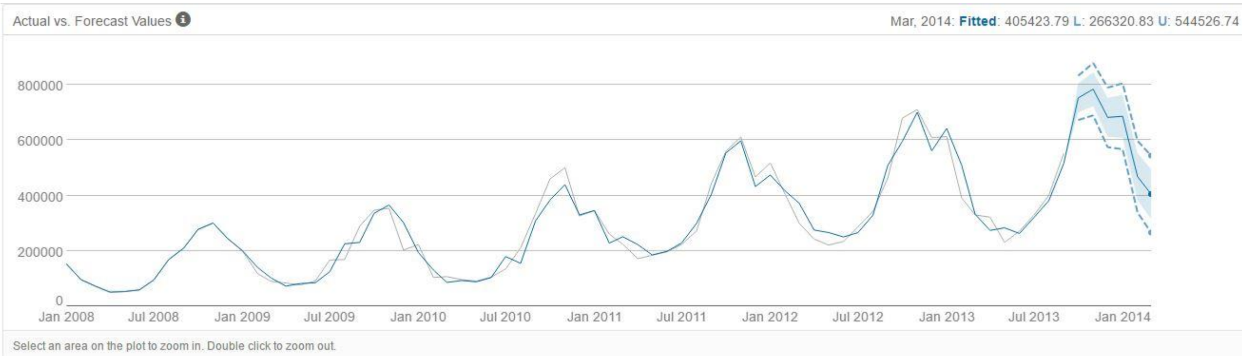
AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

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

We will be choosing the ARIMA model as it has lower AIC and MASE value.

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



The forecast for the next 6 periods(months) is shown below -

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	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
2014	2	466854.460048	594025.300958	550007.003434	383701.916662	339683.619139
2014	3	404854.460048	541411.023132	494143.997298	315564.922798	268297.896964