Analysis of Import-Export Dataset

Nishant Bhansali
Mechanical 2nd yr
IIT Roorkee

Introduction to the dataset

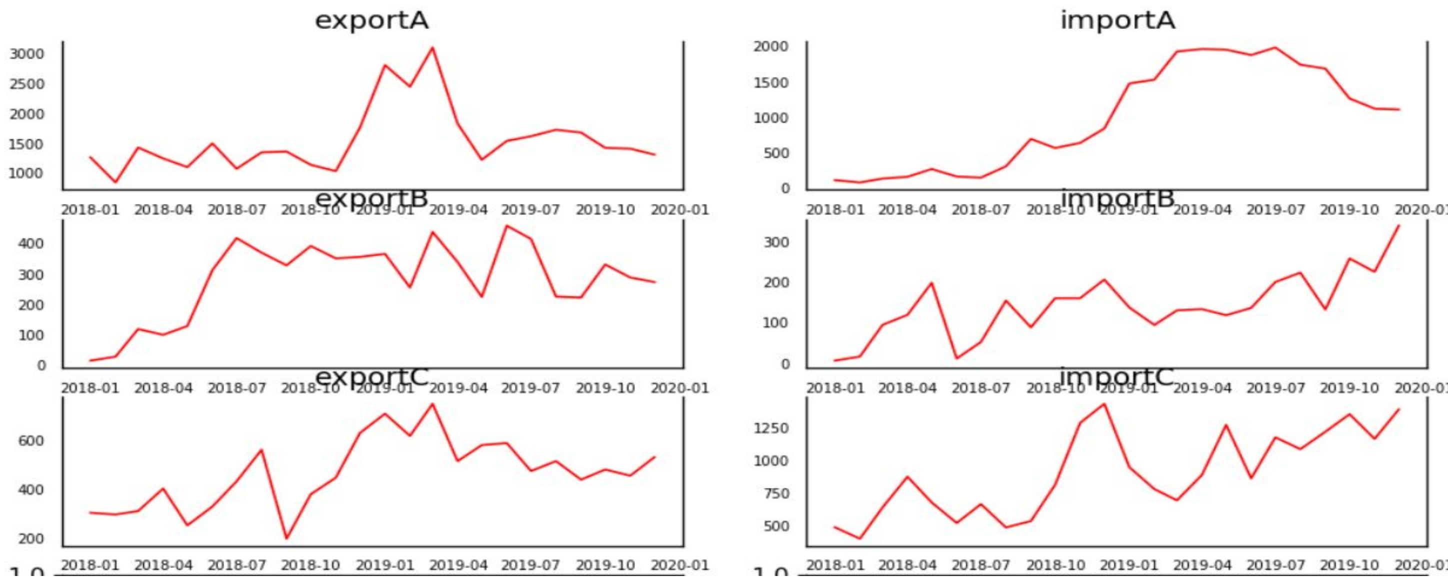
This dataset consists of information about the import and export quantities (in Tonnes) of 3 locations A,B,C.The dataset has monthly values for both the export and import of years 2018 and 2019.Datset has 24 rows and 7 columns.

month	exportA	importA	exportB	importB	exportC	importC
Jan-18	1,264	108	14	6	305	484
Feb-18	844	75	27	16	298	396
Mar-18	1,430	130	118	94	312	636
Apr-18	1,247	155	99	119	403	873
May-18	1,101	265	128	198	254	675
Jun-18	1,499	159	312	11	330	517
Jul-18	1,074	143	418	52	432	662
Aug-18	1349	303	370	154	560	483
Sep-18	1362	690	328	88	200	532
Oct-18	1140	562	392	160	380	811
Nov-18	1034	634	351	160	447	1287
Dec-18	1766	838	356	206	628	1434

Intuition behind analysis

The following dataset is a multivariate time series because we are provided with data corresponding to equal intervals of time.It is Multivariate because export and imports from different locations can affect each other,for example high export at location A may lead to low export at B or C,export at a particular location may be proportional to the import at that location.This is why to analyse this dataset I have chosen to proceed with **Vector Auto Regression** model.
It is considered as an Autoregressive model because, each variable (Time Series) is modeled as a function of the past values, that is the predictors are nothing but the lags (time delayed value) of the series.

Visualizing the data



Grangers Causality test

Using Granger's Causality Test, it's possible to test this relationship before even building the model.Granger's causality tests the null hypothesis that the coefficients of past values in the regression equation is zero.So, if the p-value obtained from the test is lesser than the significance level of 0.05, then, you can safely reject the null hypothesis.

	exportA_x	importA_x	exportB_x	importB_x	exportC_x	importC_x
exportA_y	1.0000	0.2878	0.1119	0.1562	0.0001	0.0002
importA_y	0.0955	1.0000	0.0000	0.0662	0.0000	0.0016
exportB_y	0.1939	0.1784	1.0000	0.0002	0.0471	0.0103
importB_y	0.0024	0.0002	0.3124	1.0000	0.0000	0.0195
exportC_y	0.3989	0.0001	0.0516	0.0000	1.0000	0.0000
importC_y	0.1449	0.0018	0.0014	0.0000	0.0000	1.0000

Element at index (i,j) is p-value represents the p-value of the Grangers Causality test for ith row element causing jth column element.

Cointegration test

Cointegration test helps to establish the presence of a statistically significant connection between two or more time series.Now, when you have two or more time series, and there exists a linear combination of them that has an order of integration (d) less than that of the individual series, then the collection of series is said to be cointegrated.

```
Name    ::  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
exportA ::  179.49    > 83.9383   =>   True
importA ::  93.28     > 60.0627   =>   True
exportB ::  49.93     > 40.1749   =>   True
importB ::  20.99     > 24.2761   =>   False
exportC ::  1.11      > 12.3212   =>   False
importC ::  0.12      > 4.1296    =>   False
```

From here we can see that two are more time series are definitely related so we can say that using VAR model was a good decision instead of using other models like ARIMA

Checking for stationarity

Since the VAR model requires the time series you want to forecast to be stationary, it is customary to check all the time series in the system for stationarity. If a series is found to be non-stationary, you make it stationary by differencing the series once and repeat the test again until it becomes stationary.
The results after performing the Augmented Dickey-Fuller Test (ADF Test) and performng the necessarry differences are as follows.

	INITIALLY	AFTER 1ST DIFFERENCE
EXPORT A	NON STATIONARY	NON STATIONARY
IMPORT A	STATIONARY	NON STATIONARY
EXPORT B	STATIONARY	STATIONARY
IMPORT B	NON STATIONARY	STATIONARY
EXPORT C	STATIONARY	STATIONARY
IMPORT C	NON STATIONARY	STATIONARY

2nd differnece only makes more series non stationary

Durbin Watson Static

If there is any correlation left in the residuals, then, there is some pattern in the time series that is still left to be explained by the model. In that case, the typical course of action is to either increase the order of the model or induce more predictors into the system or look for a different algorithm to model the time series.

So, checking for serial correlation is to ensure that the model is sufficiently able to explain the variances and patterns in the time series.

$$DW = \frac{\Sigma_{t=2}^{T}((e_t - e_{t-1})^2)}{\Sigma_{t=1}^{T} e_t^2}$$

The value of this statistic can vary between 0 and 4. The closer it is to the value 2, then there is no significant serial correlation. The closer to 0, there is a positive serial correlation, and the closer it is to 4 implies negative serial correlation.

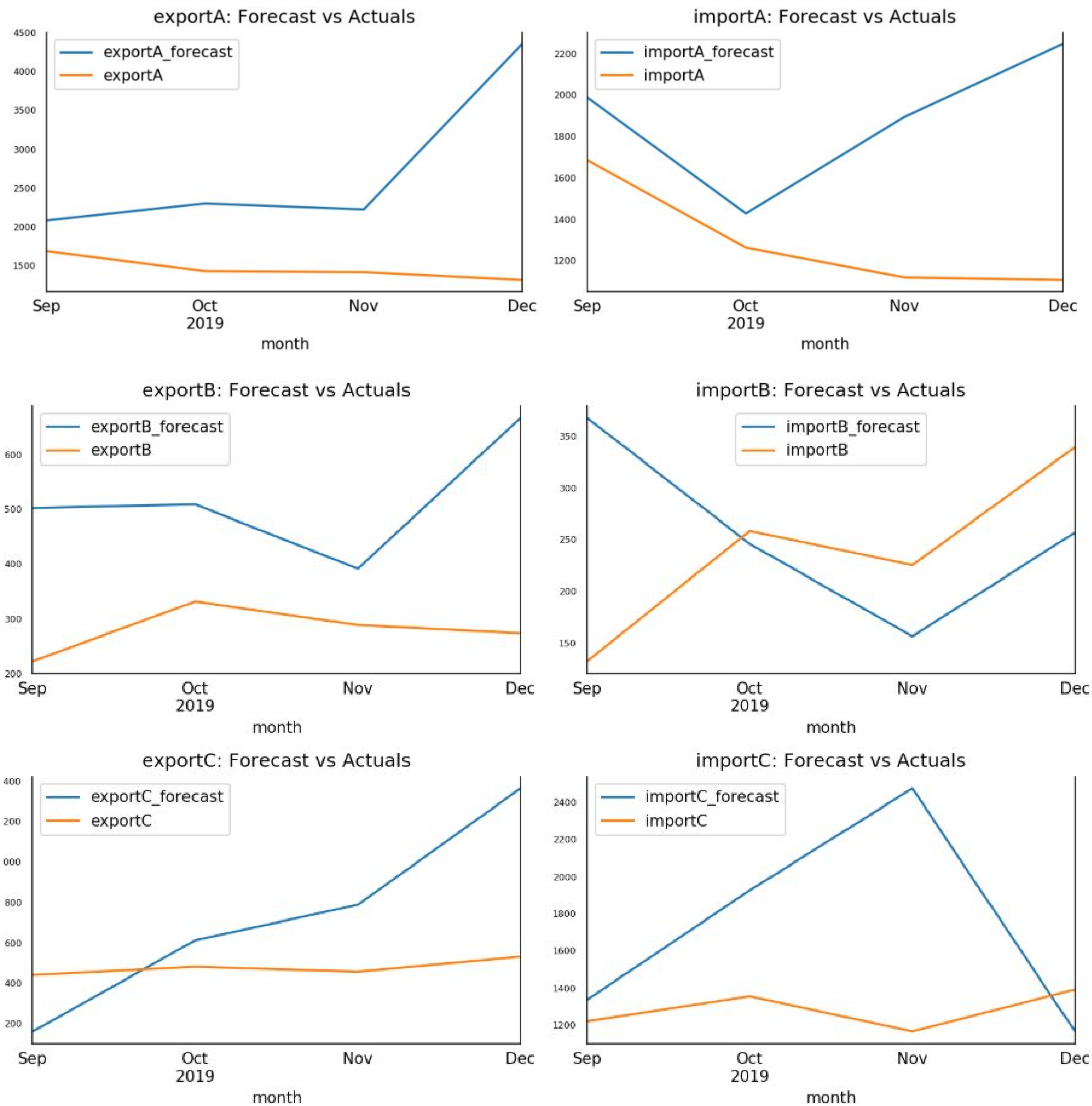
Here are the observations for the test

```
exportA : 0.86
importA : 1.12
exportB : 1.89
importB : 1.26
exportC : 0.88
importC : 1.58
```

Forecasting

In order to forecast, the VAR model expects up to the lag order number of observations from the past data.This is because, the terms in the VAR model are essentially the lags of the various time series in the dataset, so you need to provide it as many of the previous values as indicated by the lag order used by the model.Here is the forecasted data for our test set and the corresponding graphs plotted.

month	exportA_forecast	importA_forecast	exportB_forecast	importB_forecast	exportC_forecast	importC_forecast
2019-09-01	2078.739063	1986.672139	502.080901	367.208204	159.822282	1335.386805
2019-10-01	2296.289698	1426.939290	508.719140	245.370490	611.351962	1925.059840
2019-11-01	2218.875639	1893.105585	391.028452	156.030651	786.636786	2474.217784
2019-12-01	4353.571280	2245.729809	666.697439	256.504002	1362.300127	1164.969853



Metric evaluation of the forecast model

To evaluate the forecasts, let's compute a comprehensive set of metrics, namely, the MAPE, ME, MAE, MPE, RMSE, corr and minmax

MAPE :is the sum of the individual absolute errors divided by the demand (each period separately). Actually, it is the average of the percentage errors.

The Mean Absolute Error (MAE) is a very good KPI to measure forecast accuracy. As the name implies, it is the mean of the absolute error.
The Root Mean Squared Error (RMSE) is a strange KPI but a very helpful one as we will discuss later. It is defined as the square root of the average squared error.

```
Forecast Accuracy of: exportA
mape :  0.9356
me :  1280.1189
mae :  1280.1189
mpe :  0.9356
rmse :  1645.3714
corr :  -0.674
minmax :  0.4085
```

```
Forecast Accuracy of: exportB
mape :  0.8996
me :  238.6315
mae :  238.6315
mpe :  0.8996
rmse :  262.5081
corr :  -0.1054
minmax :  0.4403
```

```
Forecast Accuracy of: exportC
mape :  0.8022
me :  254.0278
mae :  393.6166
mpe :  0.4842
rmse :  473.7879
corr :  0.9031
minmax :  0.4708
```

```
Forecast Accuracy of: importA
mape :  0.5086
me :  595.6117
mae :  595.6117
mpe :  0.5086
rmse :  710.3851
corr :  -0.0538
minmax :  0.2962
```

```
Forecast Accuracy of: importB
mape :  0.5952
me :  17.7783
mae :  99.8258
mpe :  0.2958
rmse :  129.465
corr :  -0.4789
minmax :  0.3098
```

```
Forecast Accuracy of: importC
mape :  0.4513
me :  442.9086
mae :  555.9236
mpe :  0.3701
rmse :  725.847
corr :  -0.5722
minmax :  0.269
```

Conclusions

In this analysis we covered VAR from scratch beginning from the intuition behind it, causality tests, finding the optimal order of the VAR model, preparing the data for forecasting, build the model, checking for serial autocorrelation, inverting the transform to get the actual forecasts, plotting the results and computing the accuracy metrics.

To sum up,key elements about analysing this dataset was figuring out that this a different type of time series dataset in which the different time series can be dependent on each other,figuring out how and what metrics/tests to use in various stages of the analysis .This has been a great learning experience and I really enjoyed the process.