ARIMA AND SARIMA TIME SERIES MODELS ON SEASONAL DATA

AND

VAR (Vector Auto regression) ON Multivariate Dataset

Project Summary

ARIMA AND SARIMA TIME SERIES MODELS ON SEASONAL DATA

In this project the focus was mainly on how ARIMA and SARIMA Time Series models Handle seasonal data. In general, conceptually ARIMA does not handle any seasonality. In order to work with seasonal data using ARIMA, General approach to remove seasonality is to perform differencing or taking seasonal average. By taking seasonal average it will not be very helpful when seasons are very long like more than 12 months. Forecasting monthly values based on 12 months average will not give accurate results. There are also many other methods to make data stationary and remove seasonality. In this project time series assignment 4 concepts are extended with problem statement to analyze how ARIMA and SARIMA handles seasonal data. In this project Chocolate sales data from (1964 to 1967) is used. In my Approach for ARIMA Model p and q values combination is taken in such a way evaluating p and q combinations for which it has low AIC score for ARIMA Model. Through this method of obtaining p and q from lowest AIC value ARIMA (3,1,3) model was built. Every time series model should have stationary data for better and accurate predictions. Stationary time series means it has same statistical properties. The stationarity of the data is evaluated using augmented dickey fuller test. Since it's a hypothesis test based on the p values null hypothesis statement is" presence of unit root that means data is non stationary" is accepted or rejected to confirm stationarity. 0.05 level of significance is assumed. Seasonal data is provided as input for ARIMA model for predicting 20 percent of test dataset predictions RMSE is "2698.03".

SARIMA is similar to ARIMA but it also involves seasonal extra (P, D, Q m) Values that means it considers m seasonal P Lagged time values m Seasonal Q lagged error values . m is the length of the seasonality. Since Dataset has seasonality, we can use seasonal ARIMA. Based on the chocolate sales plot, PACF and ACF plots we can assume seasonal duration can be of 4 months or 12 months. Based on the plots I have implemented my idea to build to 2 SARIMA models using these two different seasons (12 months duration (P=2, Q=2) and 4 months duration (P=3, Q=3), P and Q are seasonal parameters). below are mean squared error values for both the seasons.

Sno	Model	RMSE
1	ARIMA(3,1,3)	2698.03
2	SARIMA(3,1,3,2,0,2,12)	629.91
3	SARIMA(3,1,3,3,0,3,4)	804.12

12 months seasonal duration model is best because it has low root mean squared error.

VAR (Vector Auto regression) On Multivariate Dataset

The other dataset utilised in this project is income vs spending dataset of a particular city in America from (1995 to 2015). This is multivariate dataset which involves two variables varying across time (Income and spending variables). As per the plots it seemed that there is corelation between two variables so as per my analysis from plots. I have concluded Vector auto regression model can be implemented for this data set.

VAR (Vector auto regression): it is a regression model which considers its own lagged values for regression along with other time series values. In order to build the model, data needs to be stationary. this stationarity is achieved by second order differencing for both columns in this data set. This stationarity is evaluated by augmented dickey fuller test. In Next step p is identified by analysing BIC scores of 10 lagged values of var model. at lag 3 lowest value is found. The model input is given with second order differencing so the predictions are reverted

back to original values from second order differencing. Granger causality tests were also performed for 10 lags. Model was built using p=3.The predictions for income and spending are obtained. RMSE for test data and predictions for income and spending are 43.80 and 43.35 respectively.

Sno	Predicted Variable	RMSE
1	Income	42.36
2	Spending	43.35

Based on my analysis based on granger causality test which was performed for 3 lags I came to conclusion that Chi square test result for income granger cause spending p value is greater than 0.05 at lag 3 so that means it rejects null hypothesis of "Income does not granger cause spending" and concludes there is dependency of income for predicting spending and. For spending granger cause income test p values at lag 3 is less than 0.05 accepting null hypothesis "spending does not granger cause income" so spending does not have much impact in income prediction.