Shubham Jain

2nd year, ME, IIT Roorkee

ImportExportData

Solution codes available on GitHub - https://github.com/jainshubham1120/ML-and-DL/tree/master/Time%20Series%20Prediction%20ARIMA

The dataset is based on time series analysis. This dataset is of monthly frequency.

Different Techniques of Forecasting from Time Series:

- Simple Moving Average (SMA) - The SMA is basically deal with historical data having more and more peak and valleys. Probably it would be stock data, retail data etc.
- Exponential Smoothing (SES)- This method is suitable for forecasting data with no trend or seasonal pattern
- Autoregressive Integration Moving Average (ARIMA)- It works best when your data exhibits a stable or consistent pattern over time with a minimum amount of outliers.
- Neural Network (NN) – Training a Long Short-Term Memory Neural Network (LSTM) with PyTorch for forecasting. We can use NN in any type of industry and get benefited, as it is very flexible and also doesn't requires any algorithms.
- Croston - Croston's can be seen as a specialty forecasting method that provides value in certain limited circumstance.

We'll use ARIMA model to solve this problem as it is one of the best model used in time series. Term 'Auto Regressive' in ARIMA means it is a linear regression model uses its own lags as predictors. Linear regression models, work best when the predictors are not correlated and are independent of each other. ARIMA models allow both autoregressive (AR) components as well as moving average (MA) components. You can think of the usefulness of modeling AR components as modeling the "change since last time". The usefulness of modeling MA components capture smoothed trends in the data. The (I) in ARIMA determines the level of differencing to use, which helps make the data stationary.
ARIMA models are more flexible than other statistical models such as exponential smoothing or simple linear regression. In fact, some exponential models are special cases of ARIMA models.

Steps involved in solving through ARIMA:

1. Perform exploratory data analysis - To perform exploratory analysis, let's first review the data with summary statistics and plots in Python using various packages such as matplotlib, numpy, pandas etc.

2. Decomposition of data - We will decompose the time series for estimates of trend, seasonal, and random components.

3. Test the stationarity - A stationary time series has the conditions that the mean, variance and covariance are not functions of time. In order to fit arima models, the time series is required to be stationary. We will use two methods to test the stationarity-

- Augmented Dickey-Fuller Test (ADF) - *Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical test when it comes to analysing the stationary of a series.* The test statistic > critical value, implies that the series is not stationary.
- Rolling Statistics

4. Fit a model using an automated algorithm – ARIMA

ARIMA models are among the most widely used approaches for time series forecasting. The name is an acronym for AutoRegressive Integrated Moving Average.

In an AutoRegressive model the forecasts correspond to a linear combination of past values of the variable. In a Moving Average model, the forecasts correspond to a linear combination of past forecast errors. Basically, the ARIMA models combine these two approaches. Since they require the time series to be stationary, differencing (Integrating) the time series may be a necessary step, i.e. considering the time series of the differences instead of the original one.

An ARIMA model is characterized by 3 terms: p, d, q

where,

p is the order of the AR term, It refers to the number of lags of Y to be used as predictors.

q is the order of the MA term, It refers to the number of lagged forecast errors that should go into the ARIMA

d is the number of differencing required to make the time series stationary

The value of d, therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

You can find out the required number of AR terms by inspecting the Partial Autocorrelation (PACF) plot. Partial autocorrelation can be imagined as the correlation between the series and its lag, after excluding the contributions from the intermediate lags. So, PACF sort of conveys the pure correlation between a lag and the series. That way, you will know if that lag is needed in the AR term or not.

You can look at the ACF plot for the number of MA terms. An MA term is technically, the error of the lagged forecast. The ACF tells how many MA terms are required to remove any autocorrelation in the stationarized series.

Now that you've determined the values of p, d and q, you have everything needed to fit the ARIMA model. Let's use the ARIMA implementation in statsmodel package.

5. Calculate forecasts - Finally we can plot a forecast of the time series using the forecast function, with a 95% confidence interval.

To summarize, this has been an exercise in ARIMA modeling and using various time series Python packages. It is a good basis to move on to more complicated time series datasets, models and comparisons in Python.

"It was a nice learning experience."

Thank you

Solutions:-

1. Yes, a model to forecast the import and export for the year 2020 can be built.
2. For seasonality
   On finding actual address of the given coordinates we get -
   Location A - Panipat, Haryana, 132106, Jagjivan Ram Colony Panipat India
   Location B - Shri Vajubhai Dave Marg, Ahmedabad, Gujarat, 380051, Vasna Ahmedabad India
   Location C - Hyderabad, Telangana, 500084, APHB Colony Hyderabad India