



Big Data Analytics

Dr Sirintra Vaiwsri | Email: sirintra.v@itm.kmutnb.ac.th



Time Series



Time Series Analysis

(Dietrich et al., 2015; Analyticsvidhya, 2024)

- Time series shows an ordered sequence of values over time.
- Time series analysis analyses data in a defined interval of time.
- In other words, time series analysis shows a series of orders based on time such as years, months, weeks, etc.


Time Series Analysis

(Dietrich et al., 2015; Analyticsvidhya, 2024)

- Time series analysis is useful for studying patterns in a sequence of observations.
- The analysts can identify trends, cycles, and other analyses from the time series.




Components (Dietrich et al., 2015; Analyticsvidhya, 2024)

- There are four components:
 - Trend
 - Seasonality
 - Cyclic
 - Random
- 



Trend (Dietrich et al., 2015; Analyticsvidhya, 2024)

- Trend is a long-term movement in a time series.
 - Trend is not a fixed interval of time.
 - It shows the increase or decrease of observation values over time.
 - It can be negative, positive, or null.
- 

Seasonality (Dietrich et al., 2015; Analyticsvidhya, 2024)

- Seasonality is fixed and periodic changes in the observations over time.
- It often relates to the calendar such as holidays.


Cyclic (Dietrich et al., 2015; Analyticsvidhya, 2024)

- Cyclic is periodic changed in the observations over time but it is not as much fixed as the seasonality component.
- It is uncertainty in movement and patterns.



Random/Irregularity

(Dietrich et al., 2015; Analyticsvidhya, 2024)

- 
- Random or Irregularity is unexpected situations.
 - It is often in a short time period.
 - There is some cases where the random component is needed for modelling to forecast future values.


Data Types of Time Series (Analyticsvidhya, 2024)

- Two major types:
 - Stationary:
 - Mean value should be constant in the data
 - Variance should be constant with respect to the time-frame
 - Covariance measures the relationship between two variables
 - Non-stationary means the mean-variance or covariance is changing with respect to time.



Augmented Dickey-Fuller (ADF) Test


(Analyticsvidhya, 2024)

- Augmented Dickey-Fuller (ADF) Test can be used for checking stationary.
 - It is the most popular statistical test.
 - Null Hypothesis refers to non-stationary series
 - Alternate Hypothesis refers to the stationary series
 - P-value > 0.05 ; Fail
 - P-value ≤ 0.05 ; Accept
- 





ARIMA Model (Dietrich et al., 2015; Analyticsvidhya, 2024)

- Autoregressive Integrated Moving Average (ARIMA) model is the most widely used for time series prediction
 - It is a tool for analysing time series data with the aim of getting insights and predicting trends.
- 

ARIMA Model (Dietrich et al., 2015; Analyticsvidhya, 2024)

- Parameters:
 - p - the number of lag observations
 - d - the number of times that the raw observations are differenced
 - q - the size of the moving average window

Time Series Analysis Implementation using Python:Pandas

Import libraries:

- pandas
- auto_arima and ADFTest from pmdarima.arima
- matplotlib.pyplot



Time Series Analysis Implementation using Python:Pandas



1. Import data (year_sales.csv)
 2. Use to_datetime() function to convert Year to datetime
 3. Use set_index() function as <your dataframe>.set_index('Year', inplace=True)
- Show plot

Time Series Analysis Implementation using Python:Pandas

4. Set `adf_test = ADFTest(alpha = 0.05)`
5. `adf_test.should_diff(<your dataframe>)`
6. Create train dataset from `<your dataframe>`
7. Create test dataset from `<your dataframe>`
8. Create ARIMA model using `auto_arima` function
(For parameter setting details, please find from [here](#))

Time Series Analysis Implementation using Python:Pandas

ARIMA parameter setting example:

```
auto_arima(<your train dataset>,start_p=0, d=1,  
start_q=0, max_p=5, max_d=5, max_q=5, start_P=0,  
D=1, start_Q=0, max_P=5, max_D=5, max_Q=5,  
m=12, seasonal=True, error_action='warn', trace =  
True, supress_warnings=True,stepwise = True,  
random_state=20,n_fits = 50 )
```



Time Series Analysis Implementation using Python:Pandas



9. Use summary() function to summarise the ARIMA model

10. Create prediction:

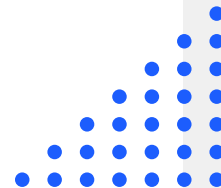
```
<your prediction> = pd.DataFrame(<your ARIMA  
model>.predict(n_periods = <number>), index = <your test  
dataset>.index)
```

```
<your prediction>.columns = ['<predicted>']
```

11. Create and show the plot that includes train, test, and predicted data.

Assignment (1 point)

- Please implement the time series analysis and show the results to get 1 point.
- The result is a plot that includes train, test, and predicted data.





References

- Dietrich, D., Heller, B., & Yang, B. (2015). Data science & big data analytics: discovering, analyzing, visualizing and presenting data. Wiley.
 - Analyticsvidhya. <https://www.analyticsvidhya.com>. Accessed: 2024-09-17.
- 