Unit 4
Time Series and Recommender Engine

Time Series Analysis

- Why Time Series Analysis?
- What is Time Series?
- Components of Time Series
- When not to use Time Series?
- What is Stationarity?
- ARIMA Model
- Case study

Why Time Series Analysis?

- In any Supervised learning,
  - Dependent and independent variable will be present and predict the function based on independent variable
- In Time Series , Analysis done on One variable i.e, time

Why Time Series Analysis?

- A very popular tool for Business Forecasting.
- Basis for understanding past behavior.
- Can forecast future activities/planning for future operations
- Evaluate current accomplishments/evaluation of performance.
- Facilitates comparison

Time Series

- An ordered sequence of values of a variable at equally spaced time intervals.
- The intervals may be hourly, weekly, monthly, quarterly, seasonally…
- *In time series, time act as an independent variable to estimate dependent variables*
- $Y= F(t)$ i.e, $Y(t) = y(t-1) + Error$
- Time Series Analysis
  - Previous behaviour
  - Plan for Future

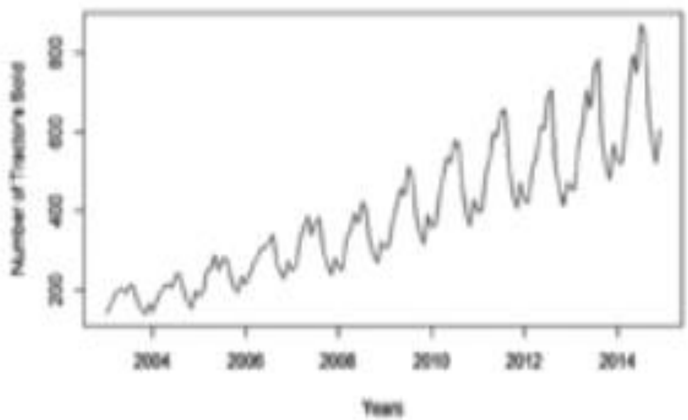
Applications

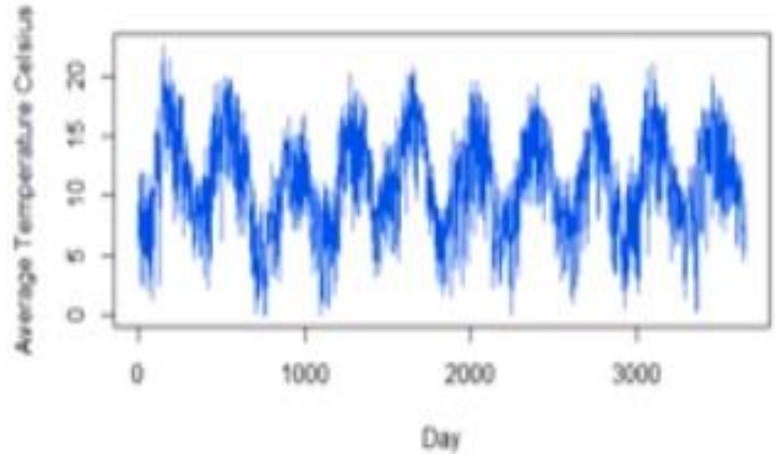
- Economic Forecasting
- Sales Forecasting
- Budgetary Analysis
- Stock Market Analysis
- Yield Projections
- Process and Quality Control
- Inventory Studies
- Workload Projections
- Utility Studies
- Census Analysis

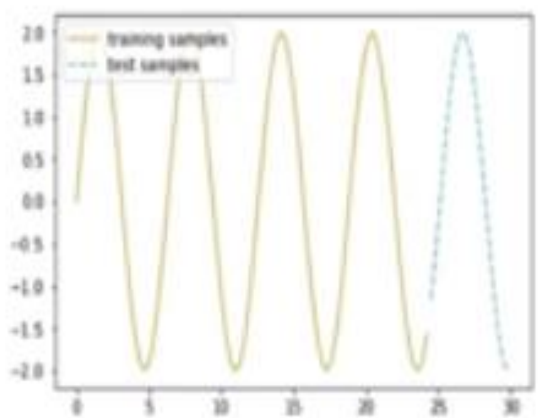
Four Components

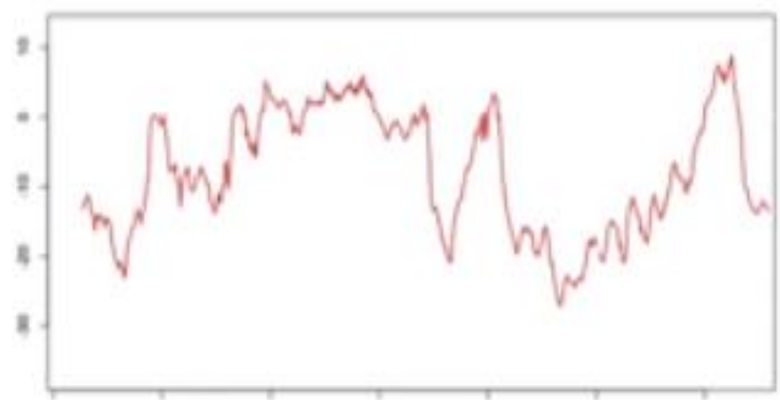
- Trend
  - Movement higher or lower for period of time
  - Happens for some time and then disappears (upward trend or downtrend or horizontal/stationary)
- Seasonality
  - It repeats itself in systematic intervals over time
- Irregularity
  - Unsystematic pattern, short duration and not repeating
  - Happens randomly
- Cyclic
  - Repeating up and down movements

Time Series data patterns









When Not to use Time Series Analysis?

- When the values are constant
- Values are in the form of functions

What is stationarity?

- A statistical property (Stationarity) always present in time series analysis
  - Constant mean (average)
  - Constant variance (distance from mean)
  - Autocovariance that does not depend on time (equal)
    - There is no correlation between the time y(t-1),y(t-2)
- Most Time series work on the assumption that TS is stationary.

Test to check stationarity

- Rolling statistics
  - Plot the moving average - We can plot the moving average or moving variance and see if it varies with time.
- Dickey Fuller Test
  - Ho: Time series is non stationary
  - Ha: Time series is stationary
  - If the ‘Test Statistic’ is less than the ‘Critical Value’, we can reject the null hypothesis and say that the series is stationary.

To make TS stationary

- Estimating and Eliminating Trend
  - Log transform
  - Moving average
  - Smoothing
  - Regression Fitting
- Eliminating Trend and Seasonality
  - Differencing
  - Decomposition

ARIMA Model

- AR
  - AR stands for autoregressive. Autoregressive parameter is denoted by p. AR terms are just lags of dependent variable. For instance if p is 5, the predictors for x(t) will be x(t-1)….x(t-5).
- MA
  - MA stands for moving the average, which is denoted by q. MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at i^{th} instant and actual value.

- I
  - In ARIMA time series analysis, Integrated is denoted by d. Integration is the inverse of differencing. When d=0, it means the series is stationary and we do not need to take the difference of it. When d=1, it means that the series is not stationary and to make it stationary, we need to take the first difference. When d=2, it means that the series has been differenced twice. Usually, more than two time difference is not reliable.

AutoRegressive(AR) Model

- Y_t depends only of past values. Y_{t-1}, Y_{t-2}, $Y_{t-3 \text{ etc}}$

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3 \ldots})$$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} \ldots$$

Moving Average Model

Y_t depends only on random error terms

$$Y_t = f(\varepsilon_{t,} \varepsilon_{t-1,} \varepsilon_{t-2,} \varepsilon_{t-3, \, ..})$$

or

$$Y_t = \beta + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3\,+...}$$

ARMA

Combines AR and MA

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} \ldots$$

$$\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3\,+\ldots}$$

Integration

- A non-stationary time series can be converted into stationary ts after differencing
- After differencing once, series is called as integrated of order 1 and denoted by I(1). In general I(d)

ARIMA Modelling

1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Recommendation Engine

Association Rule mining

- Association Rule mining is "what goes with what"
- Association rule mining is a technique to identify underlying relations between different items.
- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transactions.
- The process of identifying an associations between products is called association rule mining.
- More profit can be generated if the relationship between the items purchased in different transactions can be identified.

- For instance, if item A and B are bought together more frequently then several steps can be taken to increase the profit. For example,
  - A and B can be placed together so that when a customer buys one of the product he doesn't have to go far away to buy the other product.
  - People who buy one of the products can be targeted through an advertisement campaign to buy the other.
  - Collective discounts can be offered on these products if the customer buys both of them.
  - Both A and B can be packaged together.

- Applications
  - Market basket analysis
  - Cross-marketing
  - Catalog design etc..

Association Rules

- Association rule has to be interpreted in the form of "if-then" statements
- Association rules are probabilistic in nature

TID	ITEMS
10	Milk, Cereal, Sugar
20	Bread, Cereal, Eggs
30	Milk, Bread, Cereal, Eggs
40	Bread, Eggs

- Some possible association rules are
  - {Bread} -> {Eggs}
  - {Bread,Cereal} -> {Eggs}
- Collection of one or more items is called Itemset.

- {Bread, Cereal} -> {Eggs}

X => Y

If — Then

Antecedent - Consequent

- The possible associations can be many. We may be interested in finding the strong associations.
- But how to find strong associations ?
- Answer: Support ,Confidence & Lift.
- Support and Confidence are the measures to confirm the rule as a strong association rule.
- These two measures express the degree of uncertainty about the rule.
- The antecedent and consequent must be disjoint sets

Theory of Apriori Algorithm

- There are three major components of Apriori algorithm:
  - Support (prevalance/popularity)
  - Confidence(predictability) – likely purchase of consequent
  - Lift(interest)- association expect by chance

Three key terms to determine rules

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{\textbf{Confidence}}{\textbf{Support(Y)}}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Lift =1 means there is no association between products A and B.
Lift > 1 means products A and B are more likely to be bought together.
Lift < 1 means two products are unlikely to be bought together.

Steps in Apriori algorithm

- Find frequent itemset which satisfies the min_sup
- For each frequent itemset identify all non-empty proper subset
- For each subset s of I, form a rule s=>I where s and I are disjoint
- For each rule R, compute its confidence and Lift
- Select R as a strong rule if conf (R) >=min_conf and Lift > 1

Steps to find frequent Itemset

- Let k=1
- Generate frequent item sets of length 1
- Repeat until no new frequent item sets are identified
- **Create a candidate list** of k itemsets by performing join operation on pairs of (k-1) itemsets in the list.
- **Prune candidate item sets** containing subsets of length k that are infrequent
- **Count the support of each candidate by scanning the DB**
- **Eliminate candidates that are infrequent, leaving the list with only those that are frequent**

Example

TID	ITEMS
10	Milk, Cereal, Sugar
20	Bread, Cereal, Eggs
30	Milk, Bread, Cereal, Eggs
40	Bread, Eggs

A = milk
B= bread
C= cereal
D= sugar
E= eggs

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1[st] scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2[nd] scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3[rd] scan

L_3

Itemset	sup
{B, C, E}	2

- To speed up the process,
  - Set a minimum value for support and confidence. This means that we are only interested in finding rules for the items that have certain default existence (e.g. support) and have a minimum value for co-occurrence with other items (e.g. confidence).
  - Extract all the subsets having higher value of support than minimum threshold.
  - Select all the rules from the subsets with confidence value higher than minimum threshold.
  - Order the rules by descending order of Lift.

Advantage

- Subset of a frequent itemset is also a frequent itemset.
- This reduce the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count.
- All infrequent itemsets can be pruned if it has an infrequent subset.