Time Series Forecasting

\* We cannot use train test split in time series, we simply define that take upper 70% records for train the model and take below remaining 30% for testing the model.

\* We use only sequential data for time series.

\* Time series has a very limited scope. It is a topic that is typically only asked in interviews by companies that work on time series data or that need time series analysis. Machine learning, deep learning and SQL are more general topics that are asked in interviews for a wider range of roles.

\* **Not every problem statement requires the application of time series analysis.** Machine learning (ML) and deep learning (DL) are more general techniques that can be applied to a wider range of problems. Time series analysis is specifically designed for problems that involve time-dependent data.

\* For example: - Portfolio Management Firms, brokerage firms like Motilal Oswal, Angel Broking they requires time series.

Q) Why do we need to study time series forecasting? What are the things that machine learning (ML) and deep learning (DL) cannot solve, so we turn to time series forecasting?

A) Forecasting means predicting the future. In this, we make predictions like ML & DL, but in this, we make predictions about time series. By considering the time factor, we will do the predictions. For example, stock price. When we have to predict stock prices, it is called stock price forecasting.

\* The Indian stock market opens from 9:15 am to 3:30 pm. This duration is called trading hours. If you want to trade in the stock market, you should do an analysis of the direction in which the market is moving. Based on that, you have to predict the stock price. In predicting the stock price, two factors are important. Time factor and target variable which you want to predict.

(I) Time Factor: The time factor is the price of the stock at a particular time. For example, at 9:15 AM, the price of the stock is Rs. 200. At 10:00 AM, the price of the stock is Rs. 210. At 12:00 PM, the price of the stock is Rs. 215. And at 2:00 PM, the price of the stock is Rs. 190. Now you have to predict that what is the stock price at 3:00 pm. ow, you have to predict the stock price at 3:00 PM.

\* This is called a time series because the data depends on time. There is no third factor. There are only two factors: time and the variable you want to predict.

\* In time series, the time intervals must be equal. For example, if you are taking stock price data, make sure the data is in equal time intervals, such as 9:15, 10:15, and 11:15. The time interval difference should be 1 hour, and this is called the time interval. Similarly, we can also take a 1-day gap. For example, what is the stock price on June 1st, then June 2nd, then June 3rd? Based on this data, we need to predict the stock price on June 11th.

\* In time series, the time data should be in seconds, minutes, hours, days, weeks, etc.

\* For these types of predictions, we need time data, and to deal with time data, we need time series forecasting.

\* If you want to work in a brokerage firm, then knowledge of time series forecasting is essential.

Time Series forecasting Used In: -

(I) Stock price forecasting / prediction

(II) Weather Prediction: - Weather forecasting websites and applications use time series forecasting to predict weather. This is called **weather forecasting**. E.g.: -

|  |  |  |
| --- | --- | --- |
| Weather Forecasting | | |
| Date | Min. Temp | Max. Temp. |
| 1st June | - - |  |
| 2nd June | - - |  |
| 3rd June | - - |  |
|  |  |  |

\* In this way, we can make predictions about the minimum and maximum temperatures for each day.

\* Similarly, if the temperature data is available in hourly format, then we can predict the temperature on an hourly basis.

\* In time series, there are only two variables: the time variable and the target variable. The time variable can be in any format, such as seconds, minutes, hours, days, weeks, months, or years.

(III) Sales Forecasting: - Time series forecasting is used to predict sales. For example, companies like Nestle, Parle, and Hindustan Unilever, which manufacture any kind of products, need to create inventory in their warehouses. They do not randomly fill inventory, such as filling any type of product in any quantity. This is not the case in real life because products have expiration dates. Edible items typically have expiration dates. If they fill too many products in inventory, there is a chance that the products will expire.

\* This is where the concept of inventory management comes in. We analyse sales and then fill inventory based on that. It is especially used for products with short expiration dates, such as edible items that have expiration dates of only 1-2 months. To manage inventory, companies have a separate department that forecasts sales based on previous years' and months' data. Based on this forecast, the company produces products of each category and fills them in their inventory.

\* Generally, an error of plus or minus 10% is acceptable. However, if the error is greater than that, then they revisit the model. For example, if the monthly sales are 9 lakhs and the model predicts that sales of 10 lakhs may occur, then this is acceptable, which is a 10% error. However, if sales are only 3 lakhs and the model predicts that sales should be 10 lakhs, then this is dangerous because there is a chance that the remaining 7 lakhs of inventory will expire.

\* When we predict sales, companies give targets to their salespeople. Companies give targets that are slightly higher than the sales forecast.

(IV) Revenue Forecasting: - Customer or patient footfall refers to the number of customers or patients that are expected to visit a restaurant or hospital in the future, such as next week or next month. This information can be used to forecast revenue.

\* In all of these types of forecasting, time is considered to be an important factor.

\* News reports often state that a company is expected to generate a certain amount of revenue in the upcoming quarter. This revenue forecast is made using time series forecasting.

\* In the input data, there are many variables. However, for time series forecasting, we only need two variables. The first variable is the time variable, and the second variable is the target variable. The target variable can be the stock price, revenue, sales, or anything else. All other variables are ignored.

Q) In time series we predict the target variable and in ML we also predict the target variable. So, why we use time series instead of ML?

A) We can perform forecasting using ML and DL, but the techniques used will be different. For example, in time series forecasting, the time data is in equal time intervals along with one target variable, such as the price of something.

\* In time series forecasting, the model will consider the target variable as both the independent and dependent variable. The independent variable is the time variable, and the dependent variable is the target variable. The index is the first column in the data set.

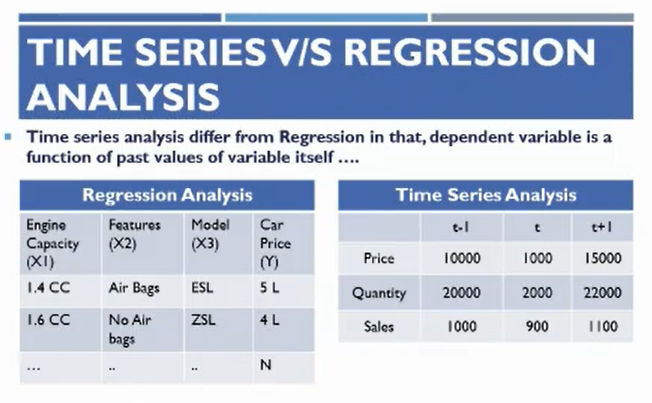
\* If the time variable is in the "object" data type, then we need to convert it into a "datetime" object. Then, we can set it as the index. After making the time variable as the index, we are left with only one variable, which is the target variable. The model will then consider the target variable as both the independent and dependent variable.

|  |  |  |
| --- | --- | --- |
| Time | | Target |
| 9:00 AM | | 100 |
| 10:00 AM | | 200 |
| 11:00 AM | | 250 |
| 12:00 AM | | 310 |
| 1:00 PM | | \_ |
|  | | |  |

\* In target variable from 1st row to 4th row (100-310) becomes predictors and price at 1:00 pm becomes target. This is called ‘Auto Regression’. This means that the model will predict the future value of the target variable based on the past values of the target variable. This is how time series forecasting works.

Q) How time series is different from regression? (Imp. I.Q)

A) The approaches of both are different. When we are working with sequential data, we need time series. In regression, we have multiple variables, but in time series, we only have two variables. In regression, we predict the dependent variable based on the independent variables. In time series, we find the future value based on the previous value. This is why sequential data is used in time series.



\* In above picture (t-1, t+1) are time lags.

\* Suppose we are working on stock price prediction, then t-1 means yesterday’s stock price, t means today’s stock price and t+1 means tomorrow’s stock price.

Sequential Data: - In sequential data, the order of the data points matters. For example, in stock price data, the price at 9:15 AM is different from the price at 10:15 AM. Time series models take the data points in sequence as input, not randomly like machine learning models.

\* This is the difference between time series and regression. In regression, the model takes the data points randomly. However, in time series, the model takes the data points in sequence.

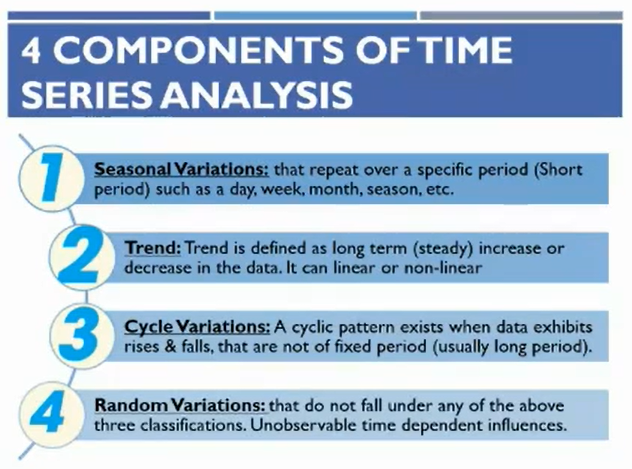
\* The same model can be built using machine learning, but it will not be as accurate because we only take the variables into account. In machine learning, we have one independent variable and one dependent variable. However, in time series, we have to consider the sequence of the data points.

\* In machine learning, we divide the data into training and test sets using train-test-validation split. This means that the data is randomly divided into two sets: the training set and the test set. However, in time series, we do not do this. Instead, we divide the data sequentially. This means that we specify that the top 70% of the data points will be used for training and the bottom 30% of the data points will be used for testing.

\* We do not use train-test-validation split in time series because it would randomly divide the data, which would break the equal time intervals and the sequence of the data points. This would break the consistency of the data, which would make it difficult to predict things accurately.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Terminologies: -

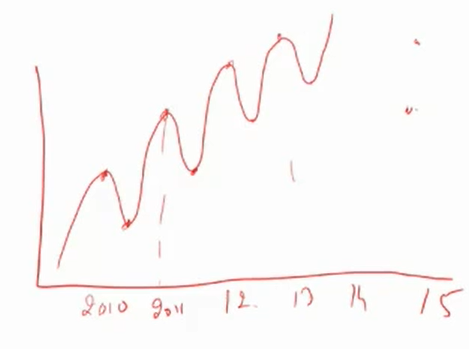


1) Seasonality: Seasonality means specific variations that repeats over a period of time. Example: - Sales of A.C is increased during summer and decrease in winters every year.

\* **Seasonality is a very important concept in time series analysis. When we perform time series analysis, we separate the seasonal component from the other components, such as the trend and the noise. This is called decomposition.**

\* Decomposition means to separate seasonality and trend component from data.

The same type of pattern is followed by the data over time, or not. For example, an upward trend. The trend is upward, and the seasonality is that the chart moves up for a certain period of time and then down for a certain period of time.



\* Ideally, the sales chart of a company looks like this, and then we say that the company is growing.

\* Seasonality matters directly or indirectly in every business.

\* For example, the sales of travel and tourism companies and soft drink companies go up for a certain period of time in a year and then go down for a certain period of time. This is where seasonality comes into play.

\* To check the performance of a company over seasonality, we check the sales of the same month in previous years and then compare them with the sales of the same month in the current year. This will perfectly indicate whether the company is growing or not.

\* When we track the performance of a company to see if it is growing or not, we consider two things: quarter-on-quarter and year-on-year. All companies announce their results on a quarterly basis. The first quarter is from April to June, and the results are announced in June. The second quarter is from July to September, and the results are announced in September. To track the performance of any company, we need to compare the results of the current quarter with the same quarter of the previous year and the sales of the current year with the sales of the previous year. Do not compare the results of the current quarter with the previous quarter when analysing any company. This is a common mistake that many people make. Always compare the results of the current quarter with the same quarter of the previous year because sales are dependent on seasonality.

2) Trends: **Trends in time series data are the long-term changes that happen over time. Example: - short term trend which is called** fluctuations**.**

**3) Cyclic Variation:** Cyclic variations are irregular changes in a time series that happen over a long period of time. They can be caused by economic, political, or technological factors. Example: Recession during covid 19.

\* In seasonality repeat variation happens over a period of time but in cyclic variation, time of variation is not fixed.

4) Random Variation: Variation which do not followed seasonality, trend and cyclic variation consider as random variation or noise.

5) Residuals

6) ARIMA

7) ARMA

8) SARIMA

9) SARIMAX

10) Stationarity

11) Ad fuller Test

12) Moving Average (Imp.): Moving average is a forecasting technique that is used to smooth out price data by averaging the price over a specified period of time.

(I) Simple Moving Average (SMA): In simple moving average, we calculate the average of the prices over a specified period of time, such as the past 200 days.

(II) Exponential Moving Average (EMA) / Weighted Average: Exponential moving average gives more weight to the most recent data than to the older data. For example, if we have data from 2010 to 2020, the exponential moving average will give more weight to the data from 2020 than to the data from 2010.

13) Time Lag: **Time lag** is the number of days of past data that are used to forecast the future. It is a parameter in time series analysis that can be defined. Time lag refers to the number of days, weeks, minutes, or any other unit of time between two observations in a time series. For example, if you are performing time series analysis on stock prices and you use data from the past month, then the time lag is 30 days.

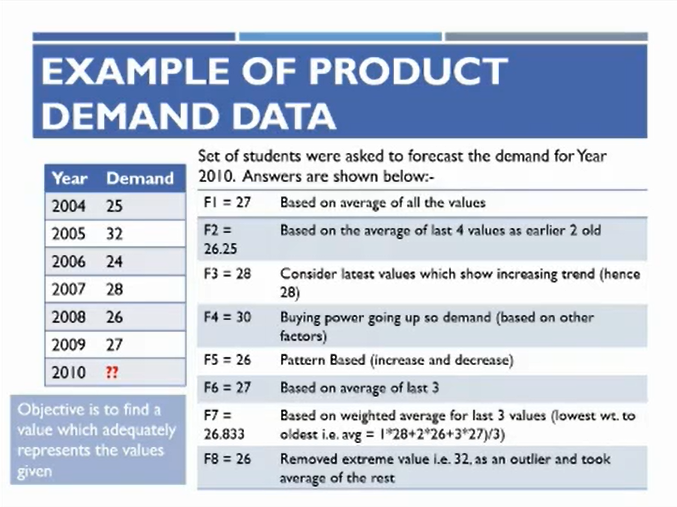
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

\* The vocabulary and dictionary of time series are totally different from what we have seen so far. Time series has its own unique terms.

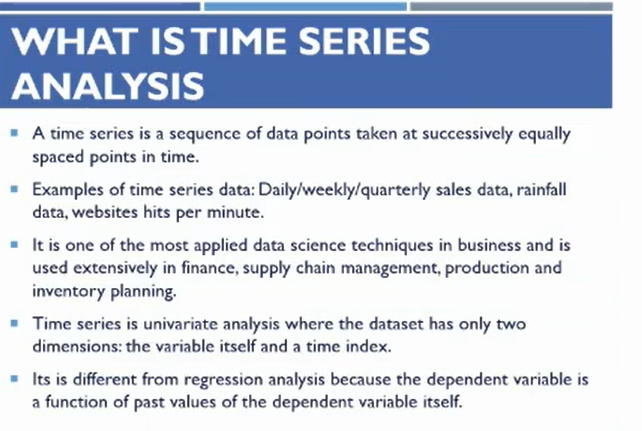
\* Time series can be used to solve problems related to time bounds. For example, if you are working in the traffic department of the Indian government and you want to make an automated traffic signal that predicts when and from which side the traffic will be heaviest. You see timers on traffic lights, which show different time durations at different times of day according to the traffic at that particular time. These timers are usually controlled manually, and you need to build an automated model that can adjust the timing of the traffic signals according to the traffic at a particular time of day. By doing this, we can reduce the waiting time of people at red lights because sometimes the traffic is heavier on one side and people have to wait for 2-3 red lights to open. To avoid this kind of problem, we can use time series. Here, time series is used because the data is in sequential form, such as traffic at 9:00 AM, then traffic at 10:00 AM, and so on.

\* If we are working in the time series department of a product company, we need to forecast quarterly sales. Based on this forecast people trade and the company forecasts their revenue.

If some not study time series than these the different ways by which he can predict the values



Q) What is time series analysis?



II Point: - At every minute what is the traffic on website.

III Point: - Above given example of inventory management of edible items.

\* In the fashion industry, they make stock according to fashion and trends. They don't produce so much that the fashion changes. Every industry has its own challenges in inventory management.

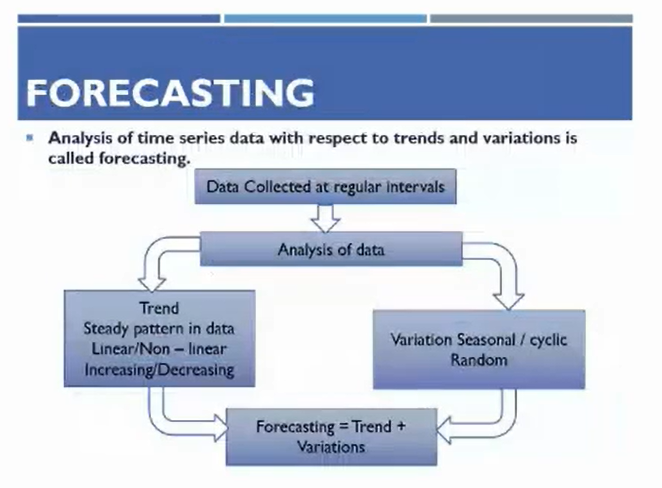
IV Point: - Above answer of how regression is different from time series.

\* When we work on time series data, we separate seasonality and trend components which is called decomposition.

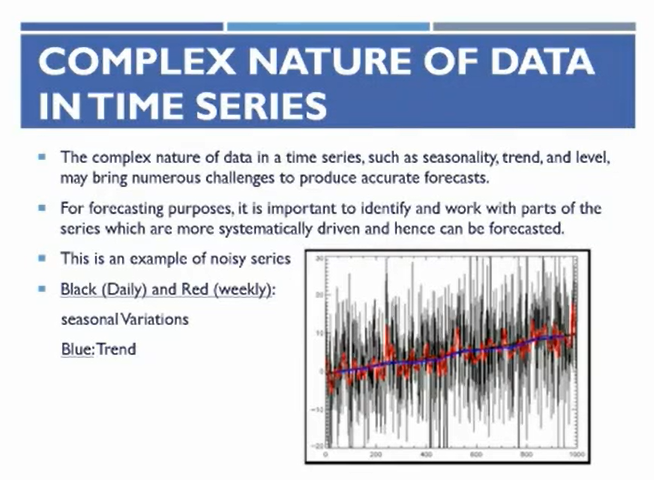
\* **Things are quite different in time series from machine learning. For example, we perform EDA differently in time series.**

**Q) What is forecasting?**

A) Analysis of time series data with respect to trends and variations is called forecasting. In ML we called it predictions and time series it is called as forecasting.

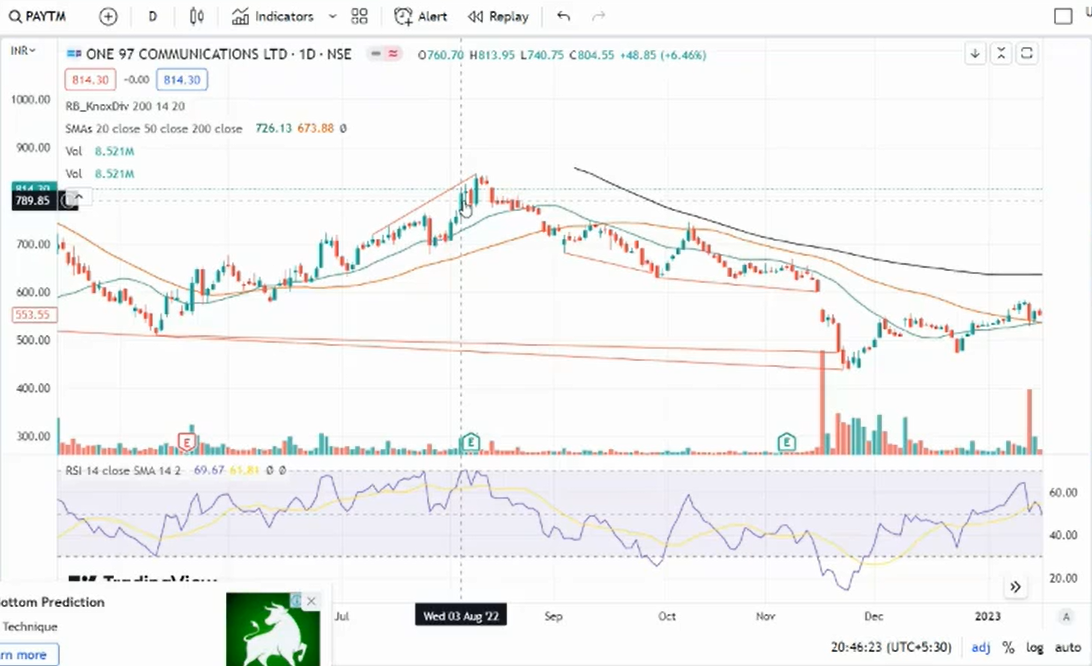
****

**Complex Nature of Data in Time Series**



**\* This is an example of noisy series where black colour indicates daily fluctuations, red colour indicates weekly fluctuations and blue indicates trend.**

**\* Complex data in time series means that there are many fluctuations in the data.**

****

**\* Stock data has 5 variables: - Open, High, Low, Close, Volume.**

**\*** The **fluctuations** between the open price and close price of a stock are called **variations / complex nature / fluctuations / Noise.** These fluctuations occur due to the trading of that particular stock on that day.

\* Volume is an important parameter when you are working with portfolio management company. When you are handling portfolio volume became an important thing.

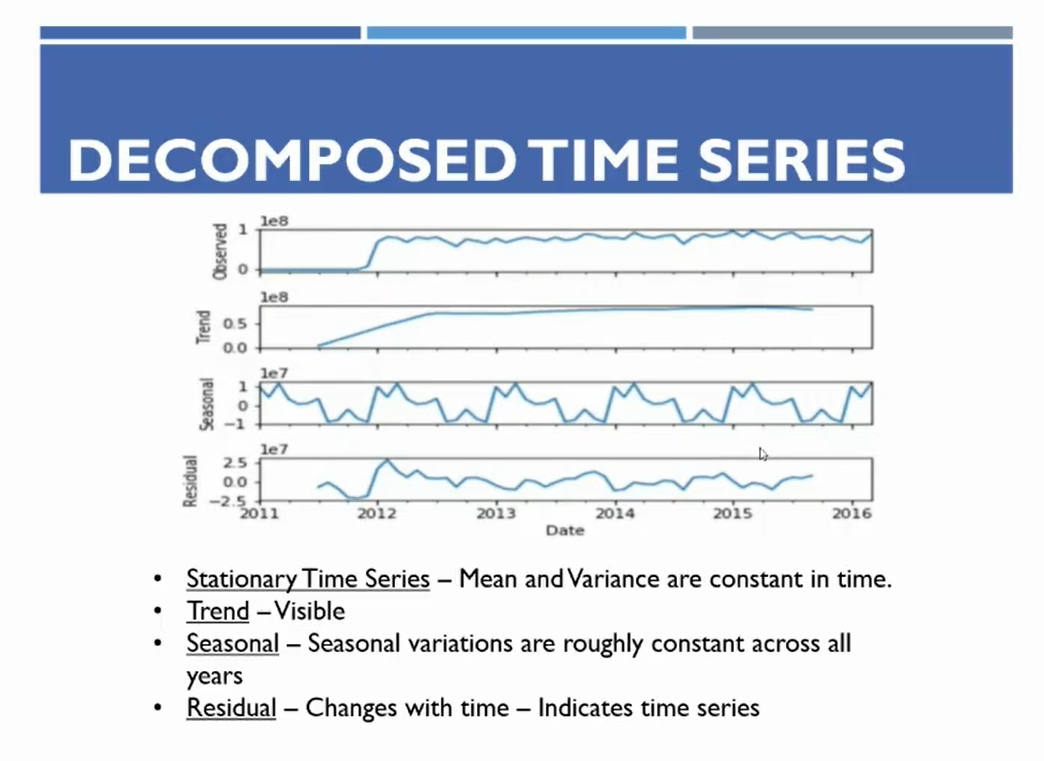
\* High trading volume indicates that many people are trading a particular stock. It is generally considered a good idea to trade stocks with high volume because it indicates that there is more liquidity in the market, which makes it easier to buy and sell the stock.

\* Low trading volume indicates that few people are trading a particular stock. This can make it difficult to buy or sell the stock, and the prices may be more volatile.

\* While working with stock data, we cannot predict all five values (open, close, high, low, and volume) simultaneously. We can only predict one value at a time. Typically, we predict the high value, which is the maximum price that the stock is expected to reach in the future. Based on this prediction, we can place a bid for the stock at that price.

\* **Bidding is the act of placing an order to buy or sell a stock at a specified price. If the stock reaches that price, the order will be executed automatically. That’s why we predict high.**

**Decomposition In Time Series**



\* In figure observed is original data and residuals are random variations / noise.

\* See that the seasonal trend goes up and then down, and this pattern repeats every year. However, the trend indicates that it is an upward trend. Therefore, it is recommended that when analysing a company's performance, you always focus on year-over-year profit or compare the same quarter's profit with the same quarter's results from the previous year.

\* **In time series analysis, when we have data with two variables, time and price, we can decompose the data into four components: seasonal variation, trend, cyclical variation, and random variation.**

**\* After building a model, we can find the error in the prediction by calculating the mean squared error (MSE). MSE is the average squared difference between the predicted values and the actual values. The lower the MSE, the better the fit of the model.**

**\*** When we have data, we can decompose it into three components: **seasonality**, **trend**, and **random variation**. **Random variation** / noise is the component of the data that is **not explained** by **seasonality** or **trend**.

**\*** For example, when the COVID-19 pandemic started, many stock prices **dropped** by 50-60% in one week. This is an example of **random fluctuation**, which **cannot be predicted**. These fluctuations **happen unpredictably** and their **timing is uncertain**. **It does not fall under seasonal variation or trend**.

\* In time series coding, there is a function called ‘seasonal\_decomposition’. This function can be used to automatically decompose the data.

Q) Why decomposition of data is required?

A) Just by looking at the data, we cannot say whether it follows an upward or downward trend, or whether it has seasonality. We can only see these things when we decompose the data into its components.

Stationarity or non- stationarity (Imp. Concept)

\* When we get data, our first task is to make it stationary. If the data is already stationary, that's great, but in real life, this doesn't happen often.



Stationarity: Stationarity means that the mean and variance of the data points are constant. In stationary data, the mean and variance should be constant at every data point or time period. This allows the model to make accurate predictions and perform better.

Non-Stationarity: Non-stationary data is data where the mean and variance are not constant. This data fluctuates over time periods, and the mean is not the same for different time periods. If we have non-stationary data, it is very difficult for the model to predict future stock prices. The model will make inaccurate predictions in this case.

\* There are two ways to check if data is stationary.

Methods by which we can identify that data is stationarity or non-stationarity

(I) The first way is to calculate the mean and variance or standard deviation of the data points. Then, check the highest and lowest values. If the difference between the highest and lowest values is more than 10%, then the data is stationary.

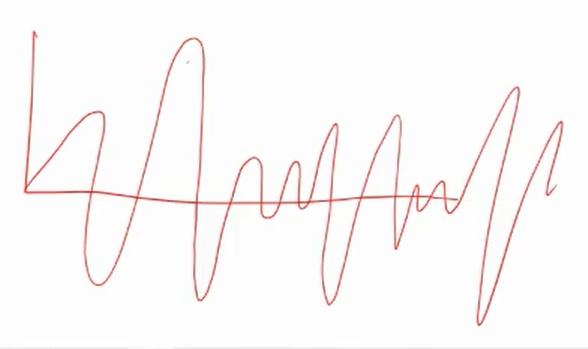
(II) Ad Fuller Test: The second way is to use the **Augmented Dickey-Fuller test** (ADF test). This is a statistical test that can be used to determine if data is stationary. If the p-value of the ADF test is less than 0.05, then the data is stationary.

Methods by which we can convert non-stationarity data into stationarity

(I) Differencing

(II) There are several transformations that can be used to make data stationary. These transformations include logarithm transformation, exponential transformation, square root transformation, and cube root transformation. After doing the transformation, we try to make the data stationary.

\* This is an example of stationary data. The length of the line is the same on the upside and the downside.



\* Stationarity in time series means that the mean and variance of the data points are constant over time. This means that the mean and variance do not change from one data point to the next.

\* When you get the data, you need to do three things: -

(I) Keep only 2 variables (Date / Target): one is the datetime variable, and the other is the target variable that you want to predict.

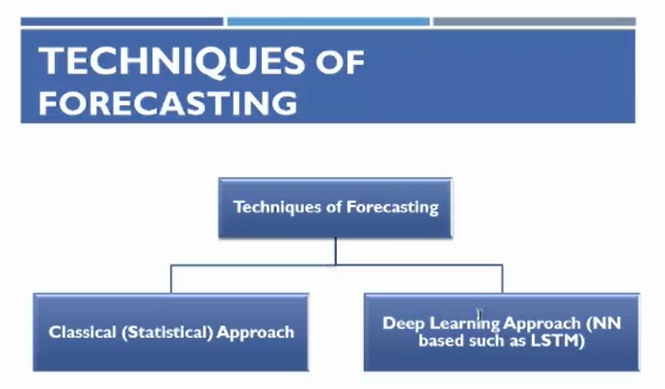
(II) If the datetime variable is in object data type, then convert it to datetime data type: This is because you will need to apply datetime functions to it, such as fetching the date, month, or year from the variable. You cannot apply these functions to object data type variables. There is a direct function available in Python for this task. The ‘to\_datetime()’ function is used to convert object data type variables to datetime data type variables.

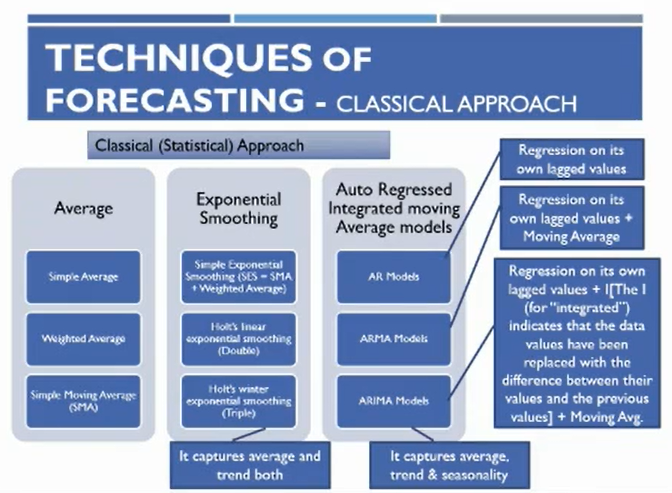
(III) Make datetime variable as index.

(IV) Make non-stationary data into stationary

\* After doing all these steps we perform other things like visualization and EDA.

Techniques Of Forecasting





\* Deep Learning has three components

(I) ANN

(II) CNN

(III) RNN: Recurrent neural networks (RNNs) are used to process sequential data, such as time series data. Here in time series, we use RNN.

\* RNNs have a major problem called short-term memory loss. This means that they can only remember a limited amount of information from the past. For example, if an RNN is trained on a sequence of 100 words, it may only be able to remember the last 10 or 20 words.

Another Example: ‘Donald trump is the president of U.S.A and he is a great man.’ In this sentence model is unable to find that he is used for whom. To avoid this type problem, we have LSTM (Long Short-Term Memory), it is an algorithm.

\* RNN has 2 advance versions which is LSTM & GRU.

LSTM: LSTM has both long term as well as short term memory’. For irrelevant words it has short term memory and for important words it has long term memory.

\* In time series we are using LSTM for forecasting.

There are 2 types of forecasting techniques: -

(I) Classical (Statistical) Approach

(a) Average: Simple Average, Weighted Average, Simple Moving Average (SMA)

(b) Exponential Smoothing

(c) ARIMA (Auto Regressed Integrated Moving Average Models)

(II) Deep Learning Approach (NN based such as LSTM)

CLASSICAL (STATISTICAL) APPROACH

(I) Moving Average (Imp.): Moving average is a forecasting technique that is used to smooth out price data by averaging the price over a specified period of time.

\* It is a very important concept in stock market analysis, time series analysis, and other forecasting applications. There are two main types of moving averages: simple moving average (SMA) and exponential moving average (EMA).

\* We are seeing Simple Moving Average and Exponential Moving Average using example on the trading view platform.

\* Average means mean moving means mean is moving. For example: - The green, red, and black lines in the picture are moving averages. The black line is the 200-day SMA, the red line is the 50-day SMA, and the green line is the 20-day SMA.

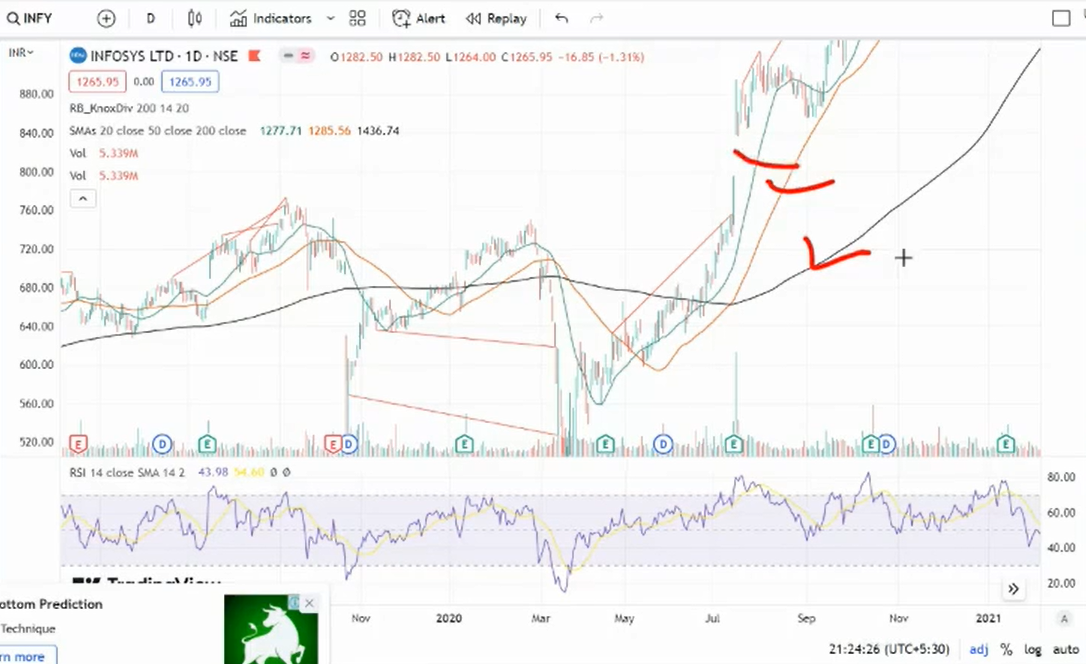
\* When we talk about moving averages, we need to decide how many days to use for the calculation. For example, a 10-day moving average, a 20-day moving average, and so on. In the stock market, the 200-day simple moving average is very popular. This is because 200 days is almost one year, since the stock market is not open on weekends. The 200-day simple moving average shows the average price of a particular stock over the past year.

Q) Why moving average is important? (Imp. Concept)

A) The 200-day simple moving average is a technical indicator that shows the average price of a stock over the past 200 days. It is often used as a long-term indicator of the trend of the stock. A common trading strategy is to buy a stock when it falls below its 200-day simple moving average and sell it when it rises above the average by 10-20%.

\* If today is June 10th, we would use the closing prices of the stock from May 1st to June 9th to calculate the 200-day simple moving average.

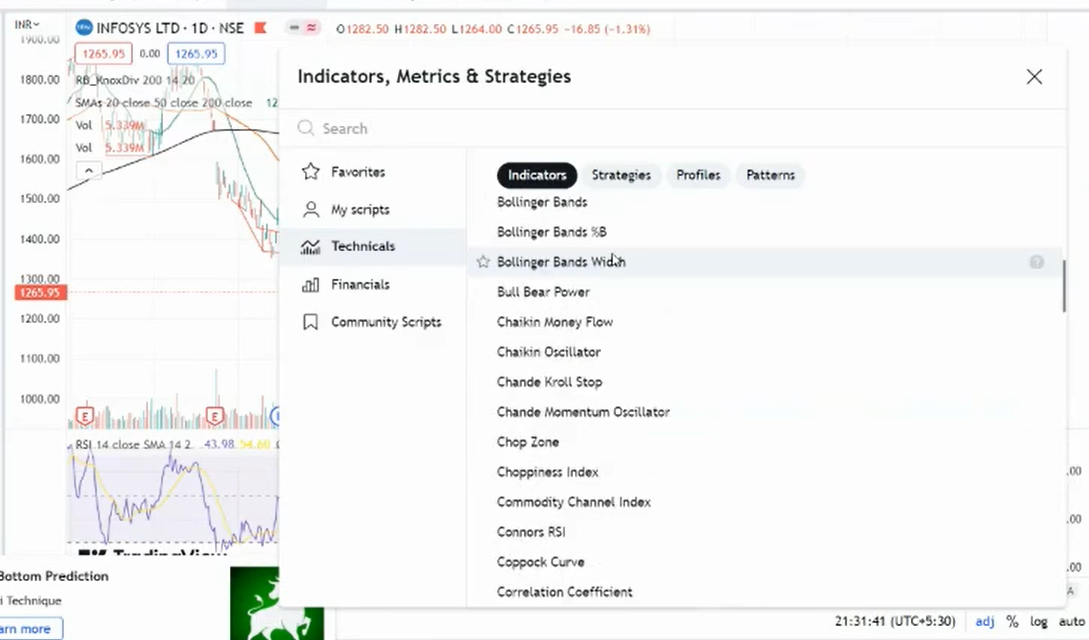
\* The number of days used in the calculation of a moving average is called the period. The period can be any number of days, but the most common periods are 10, 20, 50, and 200. As the period increases, the moving average becomes smoother. This is because the moving average is calculated using a larger number of data points, which reduces the impact of any individual data point.



(a) Simple Moving Average (SMA): In simple moving average, we calculate the average of the prices over a specified period of time, such as the past 200 days.

(b) Exponential Moving Average (EMA) / Weighted Average: Exponential moving average gives more weight to the most recent data than to the older data. For example, if we have data from 2010 to 2020, the exponential moving average will give more weight to the data from 2020 than to the data from 2010.

For example:



\* These are technical indicators that can be used to help traders make decisions about when to buy or sell stocks.

\* Example: - Pharmaceutical companies typically conduct clinical trials on animals to test the safety and effectiveness of new drugs. If the drug passes the animal trials, it is then tested on a small number of human patients. If the drug is safe and effective in the human trials, it is then tested on a larger number of patients. If the drug is approved by the FDA, it can be marketed to the public.

\* If a pharmaceutical company's drug is rejected by the FDA, the stock price of the company may decline. This is because investors may view the rejection as a negative sign for the company's future prospects.

\* The stock price of pharmaceutical companies can also be affected by FDA price capping or drug bans. For example, if the FDA caps the price of a drug, the company's profits may decrease. If a drug is banned, the company may have to destroy its inventory of the drug, which could lead to further losses.

\* The same is true for production companies. If a company's product is recalled due to safety concerns, the stock price of the company may decline. This is because investors may view the recall as a negative sign for the company's future prospects.

\* The choice of which moving average to use depends on the specific situation. In general, simple moving average is a good choice for identifying long-term trends, while exponential moving average is a good choice for identifying short-term trends.

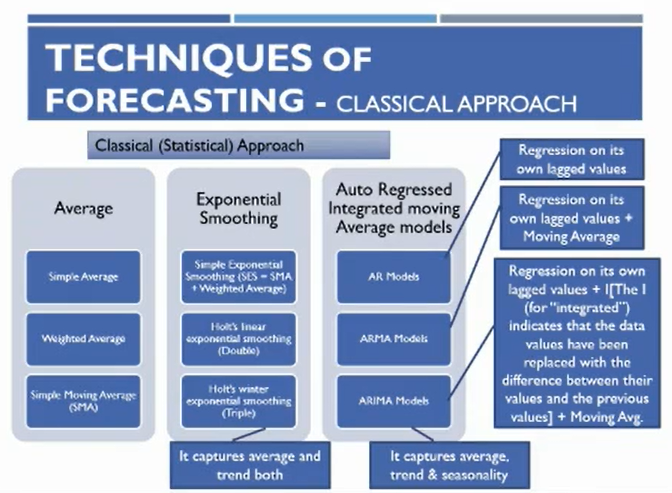
\* You have understanding that when to use which forecasting technique.

(II) Exponential Smoothing

(a) Simple Exponential Smoothing (SES – SMA + Weighted Average)

(b) Hot’s Linear Exponential Smoothing (Double)

(c) Hot’s Winter Exponential Smoothing (Triple): It captures average and trend both.



(III) ARIMA (Auto Regressed Integrated Moving Average Models)

\* ARIMA is a model that can be used to make predictions. In machine learning, we have linear regression, logistic regression, and other models. In time series analysis, we have simple moving average, exponential moving average, and ARIMA. These are all models that can be used to make predictions about future values.

(a) AR (Auto Regressed Models): It means regression on its own lagged values.

(b) ARMA (Auto Regressed Moving Average Models): It means regression on its own lagged values + Moving Average.

(c) ARIMA (Auto Regressed Integrated Moving Average Models): Regression on its own lagged values + Integration of values + Moving Average

More Forecasting Models: -

1) SARIMA (Seasonality Auto Regressed Integrated Moving Average Models)

2) SARIMAX

\* These all the types of models and it depends on us which to use.

Q) What is auto regression in ARIMA?

|  |  |
| --- | --- |
| Date | Stock Price |
| 01-Jun | 200 |
| 02-Jun | 250 |
| 03-Jun | 220 |
| 04-Jun | 280 |
| 05-Jun | - |

\* This is stock price data. We want to predict the stock price for June 5th based on the previous values. Autoregression means regression. We apply regression on the previous values of the stock price, and then these values become independent values, which are known as lag values. The number of lag values that the model takes is defined by us. This is a parameter called ‘P’, which means how many lag values we want to take for forecasting. The lag value means how many previous values are considered as input values. For example, we can take 20, 50, or any number of records. We have data for the past 10 years, but we do not need all of this data for forecasting. We can define the P parameter to take data from the past 2-3 years. This is called autoregression.

Q) What is integrated in ARIMA?

A) Integration is a process of making a time series stationary by removing the trend and seasonality. This can be done by taking the difference between consecutive values of the time series.

Integrated means to make stationarity and for this we need to do differencing of data. Differencing can be done at multiple levels, such as differencing once, twice, or three times. **Let's suppose we do differencing 1; -**

|  |  |  |
| --- | --- | --- |
| Date | Stock Price | Differencing 1 |
| 1st June | 200 | -50 |
| 2nd June | 250 | 30 |
| 3rd June | 220 | 30 |
| 4th June | 280 | -60 |
| 5th June | - | - |

\* 200-250 = -50, 250-30=30, in this way differencing 1 is performed.

\* If the data is non-stationary, we perform differencing 1. After differencing 1, we check the stationarity of the differenced data. If the differenced data is not stationary, we try to make it stationary by performing differencing 2.

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Stock Price | Differencing 1 | Differencing 2 |
| 1st June | 200 | -50 | -80 |
| 2nd June | 250 | 30 | 0 |
| 3rd June | 220 | 30 | 30 |
| 4th June | 280 |  |  |
| 5th June |  |  |  |

\* -50-30=-80, 30-30 = 0, in this way we perform differencing 2.

\* After performing differencing 2, if the data becomes stationary, we move on. Otherwise, we perform differencing 3, then 4, and so on, until the data becomes stationary.

\* This is the first technique by which we can convert non-stationarity data into stationarity data. This is called differencing technique.

\* Second method is by applying transformations.

Imp interview question: -

Q1) What is stationarity data?

Q2) Why it is important to convert non-stationarity data into stationarity?

Q3) How to check that data is stationarity or non-stationarity?

Q4) If you have non-stationarity data then how will you convert it into stationarity?

Q5) What is moving average and why it is important?

A) It indicates the direction of the average stock price movement over the past **n** number of days. This is helpful because if the stock price is below its 200-day average moving price, it is considered to be undervalued and may be a good time to buy it.

\* If stock prize is above than it’s 200 days moving average, it means stock prize is overvalued.

\* By this we can get idea of stock price by using moving average. It indicates that where to buy or sell a particular stock.