Unit-4

Predictive Analytics

Introduction:

Predictive analytics is the use of statistics and modeling techniques to forecast future outcomes. Current and historical data patterns are examined and plotted to determine the likelihood that those patterns will repeat.

Businesses use predictive analytics to fine-tune their operations and decide whether new products are worth the investment. Investors use predictive analytics to decide where to put their money. Internet retailers use predictive analytics to fine-tune purchase recommendations to their users and increase sales.

Linear Regression:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
When there is only one independent feature, it is known as **Simple Linear Regression**, and when there are more than one feature, it is known as **Multiple Linear Regression.**

Why Linear Regression is Important?

The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.
Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

Types of Linear Regression

There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y=\beta 0+\beta 1X$
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\beta nX$

where:

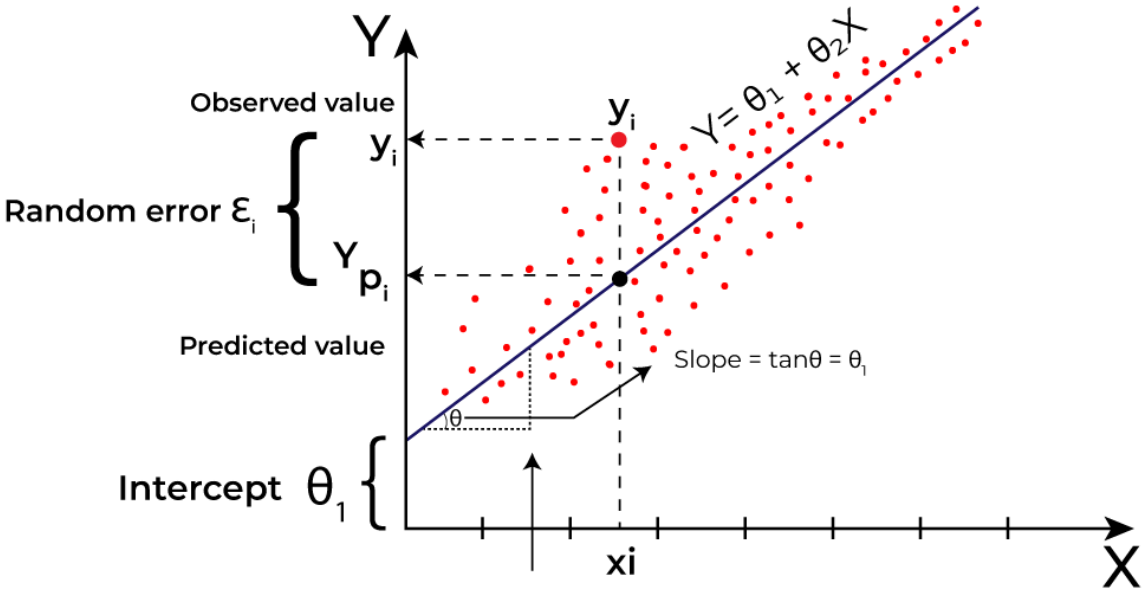
- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if we want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

What is the best Fit Line?
Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

We utilize the cost function to compute the best values in order to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.

Hypothesis function in Linear Regression

As we have assumed earlier that our independent feature is the experience i.e X and the respective salary Y is the dependent variable. Let's assume there is a linear relationship between X and Y then the salary can be predicted using:

Y=θ1+θ2X

OR
$\hat{y}_i = \theta_1 + \theta_2 x_i$

Here,

$Y_i \in Y (i=1,2,..,n)$ are labels to data (Supervised learning)
$X_i \in X (i=1,2,\ldots,n)$ are the input independent training data (univariate – one input variable(parameter))
$y_i \in \hat{Y} (i=1,2,\ldots,n)$ are the predicted values.

The model gets the best regression fit line by finding the best $\theta 1$ and $\theta 2$ values.

$\theta 1$: intercept
$\theta 2$: coefficient of x

Once we find the best $\theta 1$ and $\theta 2$ values, we get the best-fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update $\theta 1$ and $\theta 2$ values to get the best-fit line?
To achieve the best-fit regression line, the model aims to predict the target $\hat{Y}$ such that the error difference between the predicted value $\hat{Y}$ and the true value Y is minimum. So, it is very important to update the $\theta 1$ and $\theta 2$ values, to reach the best value that minimizes the error between the predicted y value (pred) and the true y value (y).

The distance between predicted value and actual value should be less. The Summation of squeare of between predicted value and actual value should be less.

$$SS = \sum_{I=1 \text{ to } n} (\hat{Y} - Yi)2$$

Regression using stats model:

statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator. The results are tested against existing statistical packages to ensure that they are correct. The package is released under the open source Modified BSD (3-clause) license. The online documentation is hosted at statsmodels.org.

Introduction

statsmodels supports specifying models using R-style formulas and pandas DataFrames. Here is a simple example using ordinary least squares:

In [1]: import numpy as np

In [2]: import statsmodels.api as sm

In [3]: import statsmodels.formula.api as smf

Multiple Regression:

Multiple regression analysis is a statistical technique that analyzes the relationship between two or more variables and uses the information to estimate the value of the dependent variables. In multiple regression, the objective is to develop a model that describes a dependent variable y to more than one independent variable.

Multiple Regression Formula

In linear regression, there is only one independent and dependent variable involved. But, in the case of multiple regression, there will be a set of independent variables that helps us to explain better or predict the dependent variable y.

The multiple regression equation is given by

$$\mathbf{y = a + b_{1\times1} + b_{2\times2} + \ldots\ldots + b_{kxk}}$$

where x_1, x_2, ….x_k are the k independent variables and y is the dependent variable.

Stepwise Multiple Regression

Stepwise regression is a step by step process that begins by developing a regression model with a single predictor variable and adds and deletes predictor variable one step at a time. Stepwise multiple regression is the method to determine a regression equation that begins with a single independent variable and add independent variables one by one. The stepwise multiple regression method is also known as the forward selection method because we begin with no independent variables and add one independent variable to the regression equation at each of the iterations. There is another method called backwards elimination method, which begins with an entire set of variables and eliminates one independent variable at each of the iterations.

Residual: The variations in the dependent variable explained by the regression model are called residual or error variation. It is also known as random error or sometimes just "error". This is a random error due to different sampling methods.

Advantages of Stepwise Multiple Regression

- Only independent variables with non zero regression coefficients are included in the regression equation.
- The changes in the multiple standard errors of estimate and the coefficient of determination are shown.
- The stepwise multiple regression is efficient in finding the regression equation with only significant regression coefficients.
- The steps involved in developing the regression equation are clear.

Logistic Regression

Logistic regression is a **supervised machine learning algorithm** used for **classification tasks** where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.

What is Logistic Regression?

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.
For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

Key Points:

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Function – Sigmoid Function

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Types of Logistic Regression

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Assumptions of Logistic Regression

We will explore the assumptions of logistic regression as understanding these assumptions is important to ensure that we are using appropriate application of the model. The assumption include:

1. Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.
2. Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.
3. Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
4. No outliers: There should be no outliers in the dataset.
5. Large sample size: The sample size is sufficiently large

Terminologies involved in Logistic Regression

Here are some common terms involved in logistic regression:

- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable's predictions.

- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.
- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

How does Logistic Regression work?

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

Let the independent input features be:

$$X=[x11.....x1m\ ,\ x21.....\ x2m,\ xn1.....xnm]$$

and the dependent variable is Y having only binary value i.e. 0 or 1.

Y={0 if class is 1,1 if class is 2}

then, apply the multi-linear function to the input variables X.

$$z = (\sum i=1 \text{ to } n\ \ wixi)+b$$

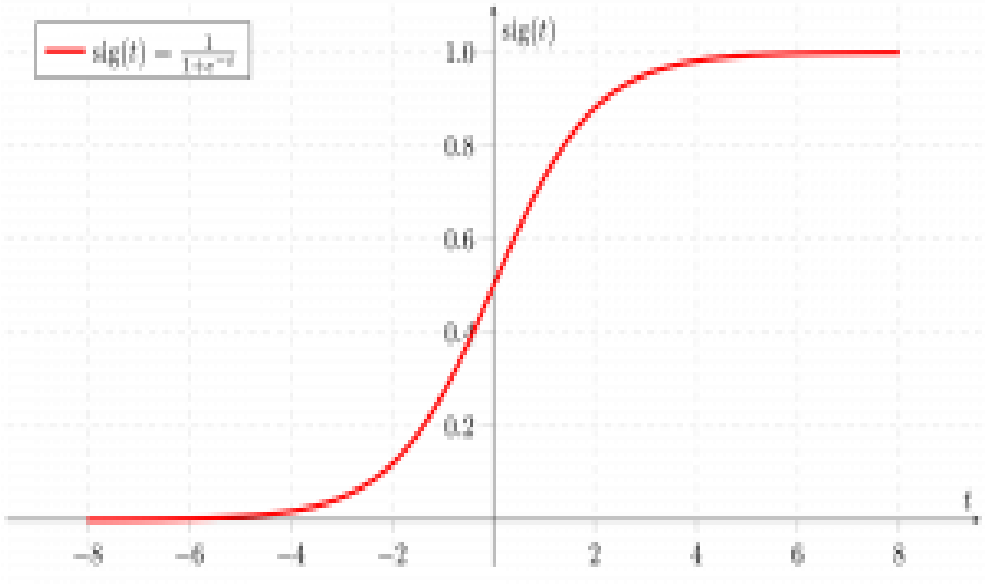
Here xi is the i th observation of X, wi= [$w1$,$w2$,$w3$,⋯,wm] is the weights or Coefficient, and b is the bias term also known as intercept. Simply this can be represented as the dot product of weight and bias.

$z=w\cdot X+b$

Sigmoid Function

Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e. predicted y.

$$\sigma(z)=1/1+e^{-z}$$



As shown above, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

- $\sigma(z)$ tends towards 1 as $z \to \infty$
- $\sigma(z)$ tends towards 0 as $z \to -\infty$
- $\sigma(z)$ is always bounded between 0 and 1

Where the probability of being a class can be measured as:

$P(y=1)=\sigma(z)$

$P(y=0)=1-\sigma(z)$

Logistic Regression-An intuitive approach

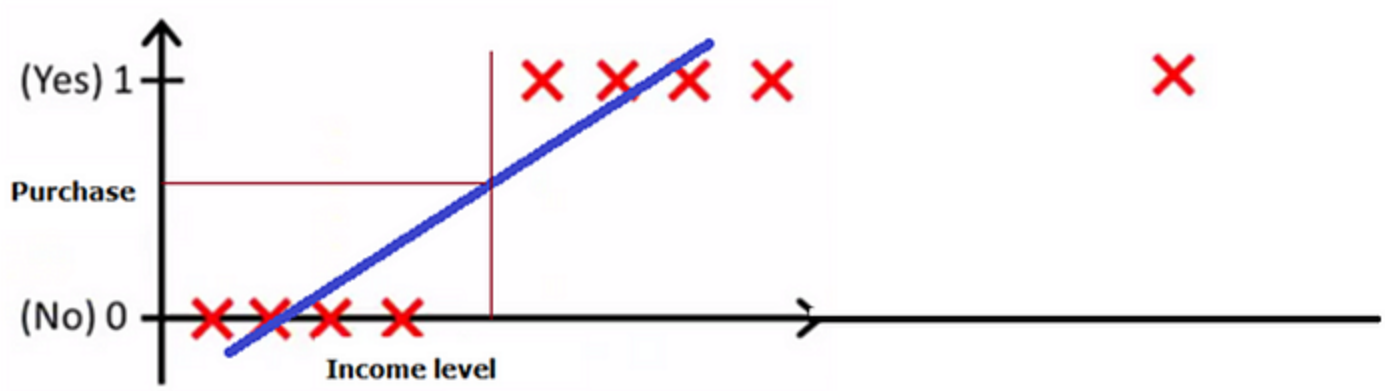
Logistic regression is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**). Logistic Regression is used when the **dependent variable(target) is categorical.**

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

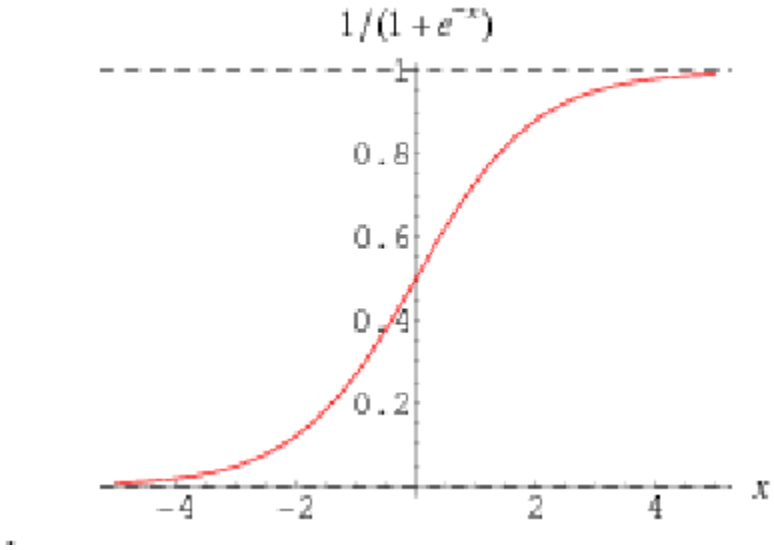
Intuition

A graph representing purchasers who either purchase as YES or NO



A graph showing that a linear line is not the ideal decision surface

In the above two graphs it is clear that a linear line is not the ideal decision surface as it may miss-classify many values , what we ideally need is a decision surface which outputs the decision as 0 or 1/yes or no.

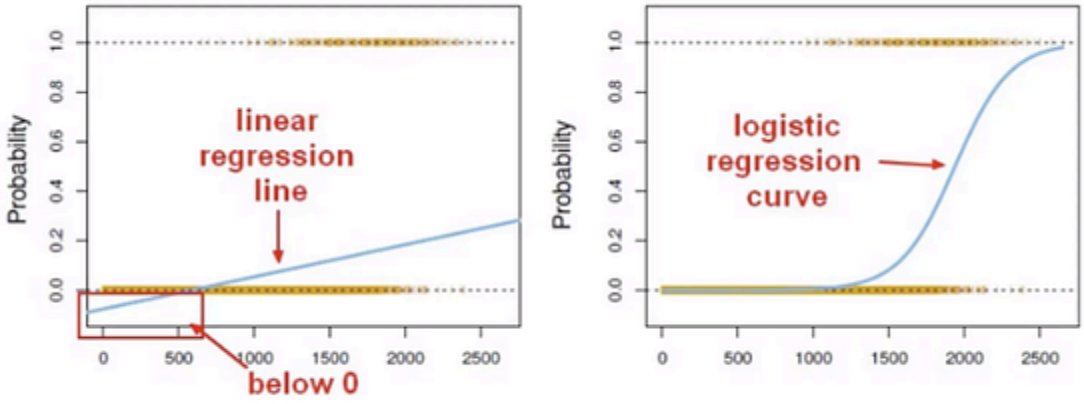


A sigmoid function

In addition, linear regression outputs values in the entire range of [-∞, ∞],whereas the actual values in this case are bound by 0 and 1.We therefore need a function that can output values between 0 and 1.A sigmoid or a logistic function allows that, and hence the name Logistic Regression.

Let us take another example

Now we have a classification problem, and we want to predict the binary output variable Y (2 values: either 1 or 0). For example, the case of flipping a coin (Head/Tail). The response is binary: 1 if the coin is Head, 0 if the coin is Tail. This is represented by a Bernoulli variable where the probabilities are bounded on both ends (they must be between 0 and 1).



The linear regression line is below 0.

Linear Regression Vs Logistic regression

Linear regression is only dealing with **continuous variables** instead of **Bernoulli variables**. The problem of Linear Regression is that these predictions are not sensible for classification since the true probability must fall between 0 and 1, but it can be larger than 1 or smaller than 0.

Note that classification is not normally distributed which is violating the **Normality assumption**. Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will not just change the mean but also the variance of the observations, which means the variance is no longer constant thus violating the assumption of **Homoscedasticity**. Thus, we cannot directly apply linear regression because it won't be a good fit.

Mathematical intuition of Logistic Regression

Before we proceed further we will acquaint ourselves with some basic mathematical terms ***Probability*** *and* ***Odds.***

The **probability** that an event will occur is the fraction of times we expect to see that event in many trials. If the probability of an event occurring is Y, then the probability of the event not occurring is 1-Y. **Probabilities** always range between 0 and 1.

The **odds** are defined as the probability that the event will occur divided by the probability that the event will not occur. Unlike **probability**, the odds are not constrained to lie between 0 and 1 but can take any value from zero to infinity.

If the probability of Success is P, then the odds of that event is:

$$odds = \frac{P}{1-P}$$

Y	*1*	*0*
Pr(Y=1)	*P*	*1- P*

**P= Success, 1-P = Failure*

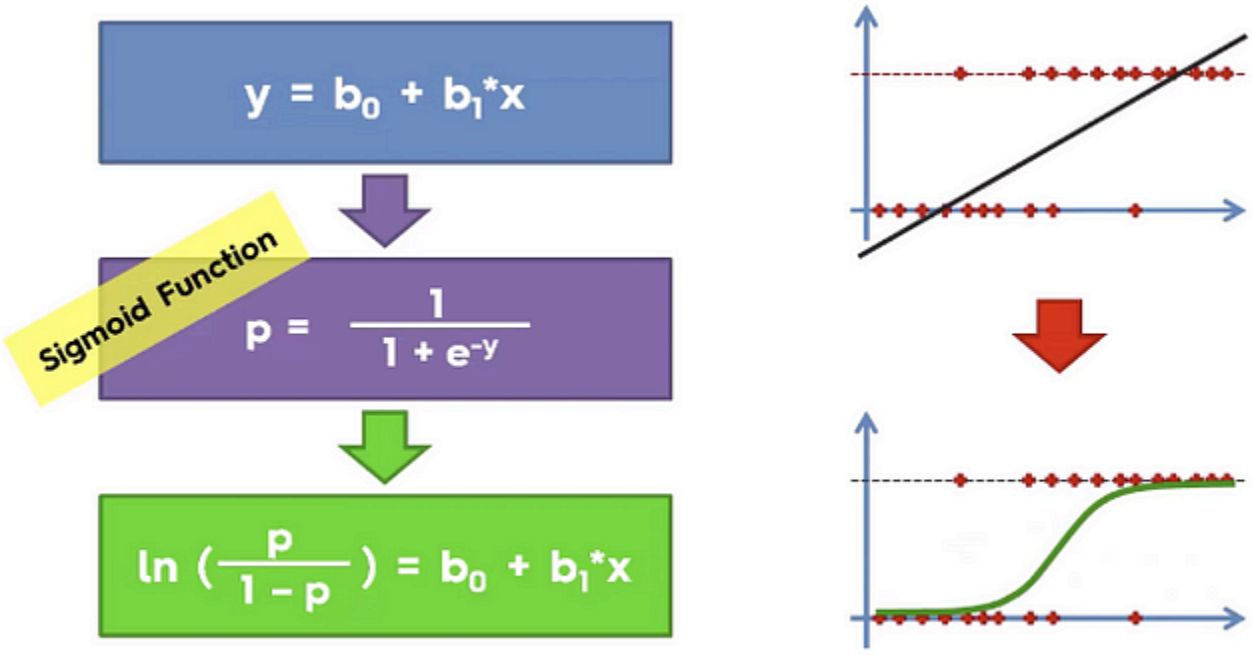
Odds: Success/ Failure.

Example: If the probability of success (P) is 0.60 (60%), then the probability of failure(1-P) is 1–0.60 = 0.40(40%). Then the odds are 0.60 / (1–0.60) = 0.60/0.40 = 1.5.

It's time… to transform the model from linear regression to logistic regression using the logistic function.

The **logit transformation** is the log of the odds ratio, that is, the log of the proportion divided by one minus the proportion. The base of the logarithm isn't critical, and e is a common base.

The logit transformation transforms a line to a logistic curve. Logistic regression fits a logistic curve to set of data where the dependent variable can only take the values 0 and 1. It can be generalized to fitting ordinal data.



logistic function (also called the 'inverse logit').

Differences between Linear and Logistic Regression

The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear regression is used for solving regression problem.	It is used for solving classification problems.
In this we predict the value of continuous variables	In this we predict values of categorical variables
In this we find best fit line.	In this we find S-Curve.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for Estimation of accuracy.
The output must be continuous value, such as price, age, etc.	Output must be categorical value such as 0 or 1, Yes or no, etc.
It required linear relationship between dependent and independent variables.	It not required linear relationship.
There may be collinearity between the independent variables.	There should not be collinearity between independent variables.

Time Series Analysis:

A time series is a sequence of data points collected, recorded, or measured at successive, evenly-spaced time intervals.

Each data point represents observations or measurements taken over time, such as stock prices, temperature readings, or sales figures. Time series data is commonly represented graphically

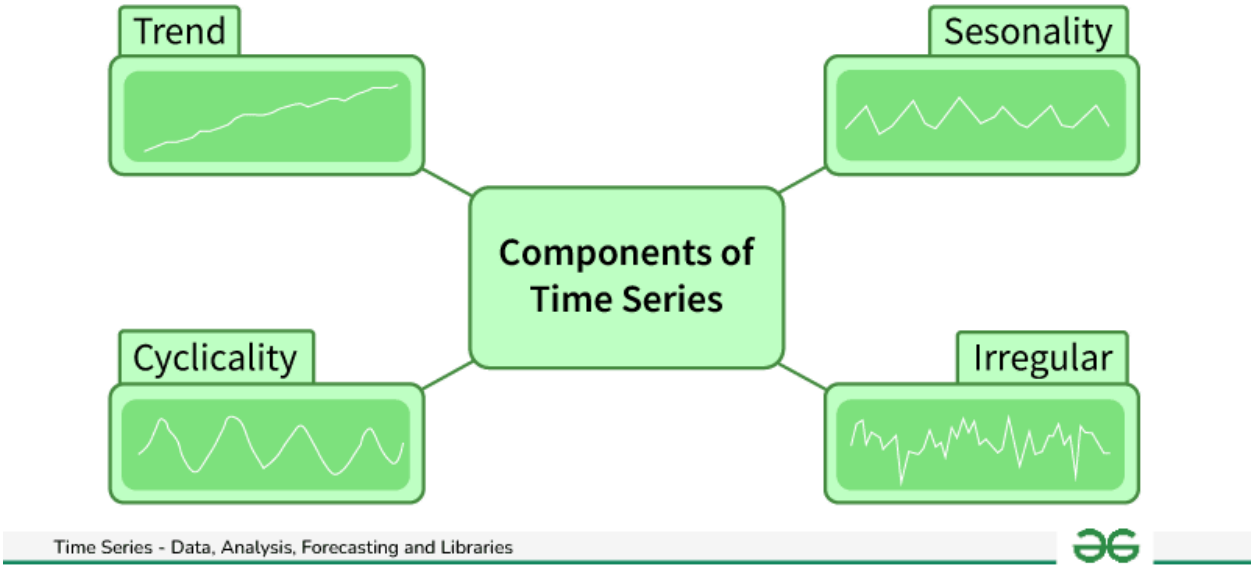
with time on the horizontal axis and the variable of interest on the vertical axis, allowing analysts to identify trends, patterns, and changes over time.

Time series data is often represented graphically as a line plot, with time depicted on the horizontal x-axis and the variable's values displayed on the vertical y-axis. This graphical representation facilitates the visualization of trends, patterns, and fluctuations in the variable over time, aiding in the analysis and interpretation of the data.

Importance of Time Series Analysis

1. **Predict Future Trends:** Time series analysis enables the prediction of future trends, allowing businesses to anticipate market demand, stock prices, and other key variables, facilitating proactive decision-making.
2. **Detect Patterns and Anomalies:** By examining sequential data points, time series analysis helps detect recurring patterns and anomalies, providing insights into underlying behaviors and potential outliers.
3. **Risk Mitigation:** By spotting potential risks, businesses can develop strategies to mitigate them, enhancing overall risk management.
4. **Strategic Planning:** Time series insights inform long-term strategic planning, guiding decision-making across finance, healthcare, and other sectors.
5. **Competitive Edge:** Time series analysis enables businesses to optimize resource allocation effectively, whether it's inventory, workforce, or financial assets. By staying ahead of market trends, responding to changes, and making data-driven decisions, businesses gain a competitive edge.

Components of Time Series Data



1. **Trend:** Trend represents the long-term movement or directionality of the data over time. It captures the overall tendency of the series to increase, decrease, or remain stable. Trends can be linear, indicating a consistent increase or decrease, or nonlinear, showing more complex patterns.
2. **Seasonality:** Seasonality refers to periodic fluctuations or patterns that occur at regular intervals within the time series. These cycles often repeat annually, quarterly, monthly, or weekly and are typically influenced by factors such as seasons, holidays, or business cycles.
3. **Cyclic variations:** Cyclical variations are longer-term fluctuations in the time series that do not have a fixed period like seasonality. These fluctuations represent economic or business cycles, which can extend over multiple years and are often associated with expansions and contractions in economic activity.
4. **Irregularity (or Noise):** Irregularity, also known as noise or randomness, refers to the unpredictable or random fluctuations in the data that cannot be attributed to the trend, seasonality, or cyclical variations. These fluctuations may result from random events, measurement errors, or other unforeseen factors. Irregularity makes it challenging to identify and model the underlying patterns in the time series data.

Time Series Visualization

Time series visualization is the graphical representation of data collected over successive time intervals. It encompasses various techniques such as line plots, seasonal subseries plots, autocorrelation plots, histograms, and interactive visualizations. These methods help analysts identify trends, patterns, and anomalies in time-dependent data for better understanding and decision-making.

Different Time series visualization graphs

1. **Line Plots:** Line plots display data points over time, allowing easy observation of trends, cycles, and fluctuations.
2. **Seasonal Plots:** These plots break down time series data into seasonal components, helping to visualize patterns within specific time periods.
3. **Histograms and Density Plots:** Shows the distribution of data values over time, providing insights into data characteristics such as skewness and kurtosis.
4. **Autocorrelation and Partial Autocorrelation Plots:** These plots visualize correlation between a time series and its lagged values, helping to identify seasonality and lagged relationships.
5. **Spectral Analysis:** Spectral analysis techniques, such as periodograms and spectrograms, visualize frequency components within time series data, useful for identifying periodicity and cyclical patterns.
6. **Decomposition Plots:** Decomposition plots break down a time series into its trend, seasonal, and residual components, aiding in understanding the underlying patterns.

These visualization techniques allow analysts to explore, interpret, and communicate insights from time series data effectively, supporting informed decision-making and forecasting.

Preprocessing Time Series Data

Time series preprocessing refers to the steps taken to clean, transform, and prepare time series data for analysis or forecasting. It involves techniques aimed at improving data quality, removing noise, handling missing values, and making the data suitable for modeling. Preprocessing tasks may include removing outliers, handling missing values through imputation, scaling or normalizing the data, detrending, deseasonalizing, and applying transformations to stabilize variance. The goal is to ensure that the time series data is in a suitable format for subsequent analysis or modeling.

- **Handling Missing Values :** Dealing with missing values in the time series data to ensure continuity and reliability in analysis.
- **Dealing with Outliers:** Identifying and addressing observations that significantly deviate from the rest of the data, which can distort analysis results.
- **Stationarity and Transformation:** Ensuring that the statistical properties of the time series, such as mean and variance, remain constant over time. Techniques like differencing, detrending, and deseasonalizing are used to achieve stationarity.

Time Series Analysis & Decomposition

Time Series Analysis and Decomposition is a systematic approach to studying sequential data collected over successive time intervals. It involves analyzing the data to understand its underlying patterns, trends, and seasonal variations, as well as decomposing the time series into

its fundamental components. This decomposition typically includes identifying and isolating elements such as trend, seasonality, and residual (error) components within the data.

Different Time Series Analysis & Decomposition Techniques

1. **Autocorrelation Analysis:** A statistical method to measure the correlation between a time series and a lagged version of itself at different time lags. It helps identify patterns and dependencies within the time series data.
2. **Partial Autocorrelation Functions (PACF)**: PACF measures the correlation between a time series and its lagged values, controlling for intermediate lags, aiding in identifying direct relationships between variables.
3. **Trend Analysis:** The process of identifying and analyzing the long-term movement or directionality of a time series. Trends can be linear, exponential, or nonlinear and are crucial for understanding underlying patterns and making forecasts.
4. **Seasonality Analysis:** Seasonality refers to periodic fluctuations or patterns that occur in a time series at fixed intervals, such as daily, weekly, or yearly. Seasonality analysis involves identifying and quantifying these recurring patterns to understand their impact on the data.
5. **Decomposition:** Decomposition separates a time series into its constituent components, typically trend, seasonality, and residual (error). This technique helps isolate and analyze each component individually, making it easier to understand and model the underlying patterns.
6. **Spectrum Analysis:** Spectrum analysis involves examining the frequency domain representation of a time series to identify dominant frequencies or periodicities. It helps detect cyclic patterns and understand the underlying periodic behavior of the data.
7. **Seasonal and Trend decomposition using Loess**: STL decomposes a time series into three components: seasonal, trend, and residual. This decomposition enables modeling and forecasting each component separately, simplifying the forecasting process.
8. **Rolling Correlation:** Rolling correlation calculates the correlation coefficient between two time series over a rolling window of observations, capturing changes in the relationship between variables over time.
9. **Cross-correlation Analysis:** Cross-correlation analysis measures the similarity between two time series by computing their correlation at different time lags. It is used to identify relationships and dependencies between different variables or time series.
10. **Box-Jenkins Method:** Box-Jenkins Method is a systematic approach for analyzing and modeling time series data. It involves identifying the appropriate autoregressive integrated moving average (ARIMA) model parameters, estimating the model, diagnosing its adequacy through residual analysis, and selecting the best-fitting model.
11. **Granger Causality Analysis:** Granger causality analysis determines whether one time series can predict future values of another time series. It helps infer causal relationships between variables in time series data, providing insights into the direction of influence.