

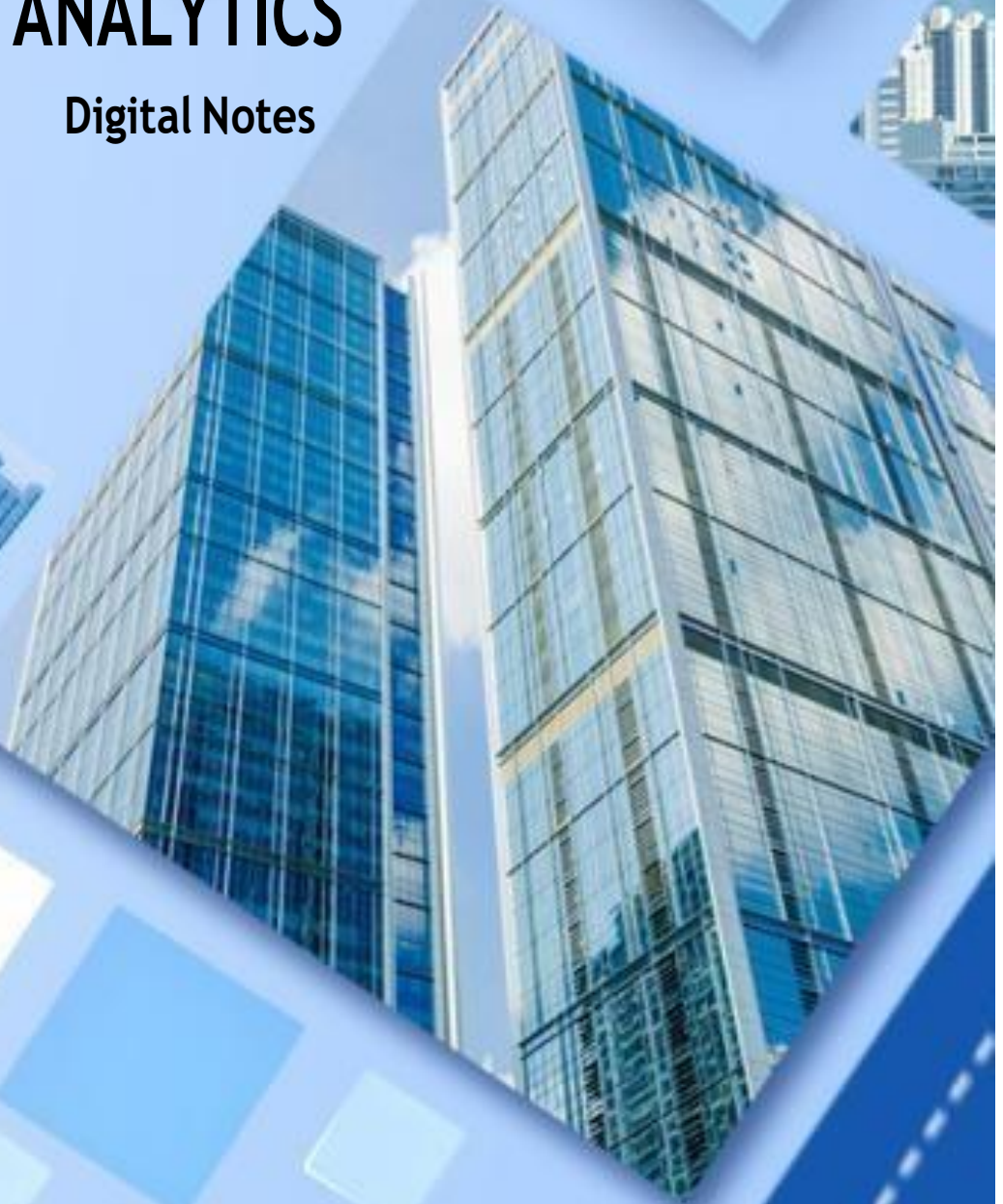


MRCET CAMPUS

ICET CODE MLRD

PREDICTIVE ANALYTICS

Digital Notes



Compiled by

DR. P. NAGA JYOTHI



SUBJECT EXPERT

Dr. P.NAGAJYOTHI
MBA, PhD

Assistant Professor,
Department of Business Management,

Malla Reddy College of Engineering & Technology.

Advice to the Students:

1. Familiarize yourself with tools commonly used in predictive analytics
2. **Regression Analysis:** Linear and logistic regression are fundamental. Understand how to apply and interpret these models.
3. **Classification and Clustering:** Be comfortable with techniques like decision trees, k-nearest neighbors, k-means clustering, and hierarchical clustering.
4. **Time Series Analysis:** Understand how to analyze and forecast time series data.
5. Solve past exam papers or sample questions to familiarize yourself with the exam format and question types.

PREDICTIVE ANALYTICS

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY (Autonomous Institution-UGC, Govt. of India)

Course : MBA II Year I SEM
Academic Year : 2023-2025
Name of the Subject : PREDICTIVE ANALYTICS
Prescribed Textbook: James R Evans, U Dinesh Kumar
Nature of the Subject: MINOR

This is an elective paper under Business Analytics specialization for MBA course, Business Analytics has become one of the most important skills that every business school student should acquire to become successful in a management career.

Course Aim: To know various predictive data analysis models and use analytical tools to solve real-life business problems. To understand basic forecasting techniques to predict future values

Learning Outcome: The students should be able to assess the suitability of Predictive models for effective business decisions.

The students will enable valid and reliable ways to collect analyze and visualize data; thereby utilize it in decision making.

To enhance the skills on linear and logistic regression.

To apply forecasting techniques in making effective business decisions

UNIT-I: Simple Regression Analysis: Concept Fundamentals of Regression Analysis - Requirements in Regression Model Building - Model Diagnostics - Interpretation of Regression results for Management Decision.

Multiple Regression Analysis: Concept - Significance of Multiple Regression Analysis - Structure of Model Estimation - Testing Rule of Multiple Regression Analysis **Unit-II: Non-linear Regression and Regression Modeling**

UNITII: Non-Linear Regression Analysis: Concept - Types of Non-linear Regression Models - Model Transformation - Difference between Linear and Non-linear Regression Models.

Diagnostics of Regression Modelling: Model Diagnostics - Multicollinearity - Autocorrelation

Unit-III: Dummy modelling and Panel Data Model Dummy modeling: Dummy independent modelling-linear probability Model-Logit model-Probit model

Panel Data Model: Concept - Panel Data Models - Fixed Effects Model - Random Effects Model - Forms of Panel Data Models - Applications to use Panel Data Models.

PREDICTIVE ANALYTICS

Unit-IV: Forecasting and Machine Learning: Time Series Forecasting: Concept - Forecasting Techniques - Measures of Forecast Error - Trend Analysis - Time Series Models - Auto Regressive Model - Applications of Time Series Models.

Machine Learning: Concept - Predictive Analysis under Machine Learning - Model of Artificial Neural Networks (ANN) - Model of Random Forest - Model of Support Vector Machine - Assumptions under Machine Learning.

Unit-V: Data Mining and Simulation: Data Mining: Concept - Data Interpretation - Data Reduction - Classification and Clustering Techniques - Association Rule Mining - Cause and Effect Model.

Simulation: Concept - Monte Carlo Simulation - Discriminant Event Simulation - Application Using Simulation.

PREDICTIVE ANALYTICS

UNIT-1

Predictive analytics has consistently increased over the last five years. Predictive analytics (also known as advanced analytics) is increasingly being linked to [business intelligence](#).

Understanding Predictive Analytics

Let us take an example of a certain organization that wants to know what its profit will be after a few years in the business, given the current trends in sales, the customer base in different locations, etc. Predictive analytics will use the variables given, and techniques such as data mining and artificial intelligence will predict the future profit or any other factor the organization is interested in.

What is Predictive Analytics?

Predictive analytics is a significant analytical approach used by many firms to assess risk, forecast future business trends, and predict when maintenance is required. [Data scientists](#) use historical data as their source and utilize various [regression models](#) and [machine learning techniques](#) to detect patterns and trends in the data.

Here are a few examples of how businesses are using predictive analytics:

Customer Service

Businesses may better estimate demand by utilizing advanced and effective analytics and business intelligence. Consider a hotel company that wants to estimate how many people will stay in a certain area this weekend so that they can guarantee they have adequate employees and resources to meet demand.

Higher Education

Predictive analytics applications in higher education include enrollment management, fundraising, recruiting, and retention. Predictive analytics offers a significant advantage in each of these areas by offering intelligent insights that would otherwise be neglected.

A prediction algorithm can rate each student and tell administrators ways to serve students during the duration of their enrollment using data from a student's high school years.

Models can give crucial information to fundraisers regarding the optimal times and strategies for reaching out to prospective and current donors.

Supply Chain

Forecasting is an important concern in manufacturing because it guarantees that resources in a supply chain are used optimally. Inventory management and the shop floor, for example, critical spokes of the supply chain wheel that require accurate forecasts to function.

Predictive modeling is frequently used to clean and improve the data utilized for such estimates. Modeling guarantees that additional data, including data from customer-facing activities, may be consumed by the system, resulting in a more accurate prediction.

Insurance

PREDICTIVE ANALYTICS

Insurance firms evaluate policy applicants to assess the chance of having to pay out for a future claim based on the existing risk pool of comparable policyholders, as well as previous occurrences that resulted in payments. Actuaries frequently utilize models that compare attributes to data about previous policyholders and claims.

Software Testing

Predictive analytics can help you enhance your operations throughout the full [software testing life cycle](#).

Simplify the process of interpreting massive volumes of data generated during software testing by using that data to model outcomes. You can keep your release schedule on track by monitoring timelines and utilizing predictive modeling to estimate how delays will affect the project. By identifying these difficulties and their causes, you will be able to make course corrections in individual areas before the entire project is delayed.

Predictive analytics can assess your clients' moods by researching social media and spotting trends, allowing you to anticipate any reaction before it occurs.

So far we discussed what is Predictive analytics and its examples. Moving forward, let's understand what are its analytics tools.

Applications of Predictive Analytics: -

1. Marketing: -

- Predictive analytics tools could be helpful in segmenting the marketing leads by displaying ads over websites and social media platforms relating to consumer behaviour and interest.
- Predictive analytics tools can explore “expect to purchase” by analysing consumer’s behaviour on past and current available data to find people whose data matches with ideal consumers.
- Marketers could also use predictive analytics for leads scoring by analysing data to identify which prospects are potentially most valuable for the company

2. Retail

1. Customer sales data provides personalized recommendations and promotions for individual customers, through predictive analytics, better targeting built over real-time data assists retailers for planning campaigns, making ads and promotions that buyers will respond the most.
2. Sales and promotion timing has become an art, conducting predictive analytics over customer, inventory, and historical sales data provides suitable circumstances/timing for lowering or raising prices.
3. Predictive analytics lets retailers in merchandise planning and price optimization to investigate the impact of promotional events and to figure out appropriate offers for consumers.

PREDICTIVE ANALYTICS

3. Manufacturing: -

1. Predictive analytics is helpful when combined with machine data in order to help in tracking and comparing machines' performance and equipment maintenance status and predicting which particular machine will fail.
2. Predictive analytics insights can lead to decrease in shipping and transportation expenses by accepting all the factors included in transferring manufacturing products at different places under the proper system.
3. Considering predictions over supply chain and sales data helps in making more considerable decisions for purchasing and ensuring that no expensive raw materials get purchased unless not required. This data can also be used in aligning manufacturing processes with consumer demands.

4. Finance: -

1. Prohibition of credit card fraud via indicating unusual transactions,
2. Credit card scoring to determine whether to approve or deny loan applications,
3. Most importantly, analyzing customers' churn data and facilitating banks to approach potential customers before they are likely to switch respective institutions.
4. Measuring credit risk, maximizing cross-sell/up-sell opportunities and retaining valuable customers.

5. Healthcare: -

1. Predictive analytics can help medical practitioners by analyzing data concerning global diseases [statistics](#), drug interactions, patient diagnostic history individually to provide advanced care and conduct more effective medical practices.
2. Applying predictive analytics on clinics' past appointment data helps in identifying probable no-shows or delays in cancellations more accurately and thus save time and resources.

To detect claims frauds, the health insurance industry is using predictive analytics to discover patients at most risk of incurable or chronic disease, it helps companies in finding suitable interventions.

CONCEPT OF ASSOCIATION

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of [machine learning](#), and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here

PREDICTIVE ANALYTICS

market basket analysis is a technique used by the various big retailers to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:



Association rule learning can be divided into three types of algorithms:

Apriori

Eclat

F-P Growth Algorithm

How does Association Rule Learning work?

Association rule learning works on the concept of If and Else Statement, such as if A then B.



Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known as *single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

Support

Confidence

Lift

PREDICTIVE ANALYTICS

The basic technique aims to find the relationship and establish the patterns on the items purchased. May in much more simpler terms, we can compare to a If-then clause, eg., If bread is bought then possibility to buy jam/butter/milk/eggs.

IF [bread] then [Jam/Butter/Milk/]

In the shorthand notation, which translates to “the items on the right are likely to be ordered with the items on the left”

Key Terms to be known in Market Basket Analysis

- Item Set: Collection of items purchased by customer.
- Antecedent: The items on the LEFT i.e., the item which the customer buy
- Consequent: The items on the RIGHT i.e., the item which the customer follows to buy.
- Support: The probability that the antecedent event will occur i.e., the customer will buy bread.
- Confidence: The probability that the consequent will occur wrt the given antecedent. i.e., the customer will buy jam/butter/milk/eggs only when he buys bread
- Lift: The lift of the rule is the ratio of the support of the left-hand side of the rule (sandwich) co-occurring with the right-hand side (tea), divided by the probability that the left-hand side and right-hand side co-occur if the two are independent.

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the item set X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{Freq}(X)}{T}$$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

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

Lift

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{\text{Supp}(X,Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

If Lift= 1: The probability of occurrence of antecedent and consequent is independent of each other.

Lift>1: It determines the degree to which the two item sets are dependent to each other.

Lift<1: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

- A lift greater than 1 suggests that the presence of the antecedent increases the chances that the consequent will occur in a given transaction
- Lift below 1 indicates that purchasing the antecedent reduces the chances of purchasing the consequent in the same transaction. Note: This could indicate that the items are seen by customers as alternatives to each other
- When the lift is 1, then purchasing the antecedent makes no difference on the chances of purchasing the consequent

An example of Association Rules

- Assume there are 100 customers
- 10 of them bought milk, 8 bought butter and 6 bought both of them.
- bought milk => bought butter
- support = P(Milk & Butter) = 6/100 = 0.06
- confidence = support/P(Butter) = 0.06/0.08 = 0.75
- lift = confidence/P(Milk) = 0.75/0.10 = 7.5

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

Market Basket Analysis: It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.

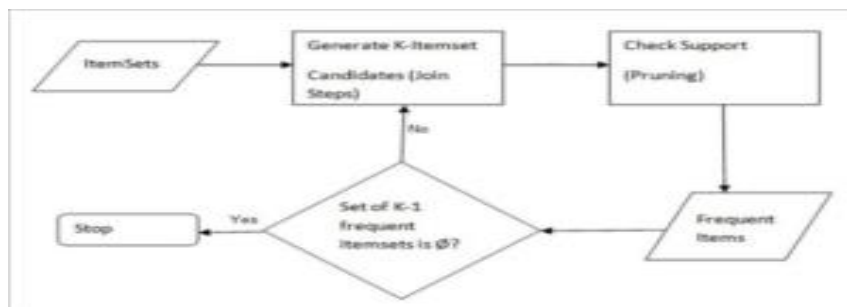
Medical Diagnosis: With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.

Protein Sequence: The association rules help in determining the synthesis of artificial Proteins. It is also used for the **Catalog Design** and many more other applications.

Algorithms of Associate Rule In Data Mining

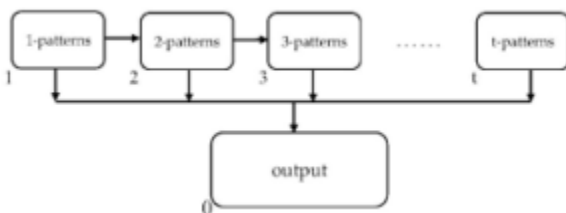
Apriori Algorithm

Apriori algorithm is a standard algorithm in Association Rule Learning in Data mining. It is used for drawing familiar item sets and their relevant association rules. It is designed to perform on a database.



Eclat Algorithm

Eclat algorithm is a type of Association rule learning algorithm. It can be applied to achieve itemset mining. Itemset mining helps us to obtain periodic patterns in data. For example, if a consumer buys shoes, he would also buy socks.

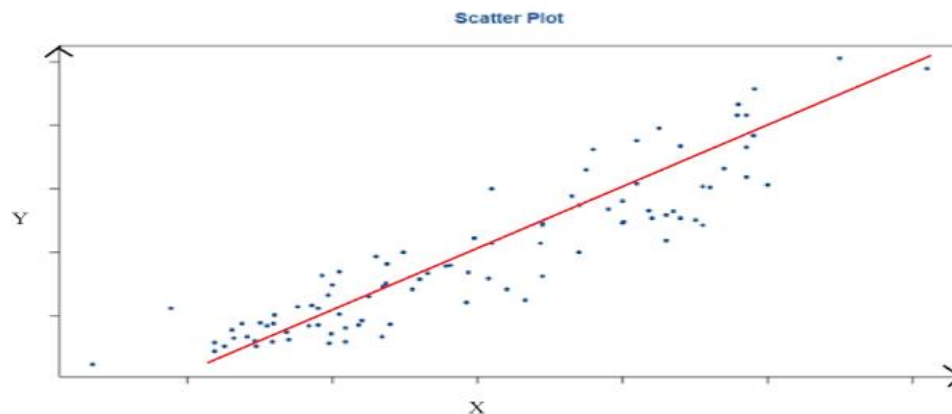


PREDICTIVE ANALYTICS

Simple Linear Regression(Interpretation of regression results for management decision)

Simple linear regression is used to find out the best relationship between a single input variable (predictor, independent variable, input feature, input parameter) & output variable (predicted, dependent variable, output feature, output parameter) provided that both variables are continuous in nature. This relationship represents how an input variable is related to the output variable and how it is represented by a straight line.

To understand this concept, let us have a look at scatter plots. Scatter diagrams or plots provides a graphical representation of the relationship of two continuous variables.



After looking at scatter plot we can understand:

1. The direction
2. The strength
3. The linearity

The above characteristics are between variable Y and variable X. The above scatter plot shows us that variable Y and variable X possess a strong positive linear relationship. Hence, we can project a straight line which can define the data in the most accurate way possible. If the relationship between variable X and variable Y is strong and linear, then we conclude that particular independent variable X is the effective input variable to predict dependent variable Y.

To check the co-linearity between variable X and variable Y, we have correlation coefficient (r), which will give you numerical value of correlation between two variables. You can have strong, moderate or weak correlation between two variables. Higher the value of “r”, higher the preference given for particular input variable X for predicting output variable Y. Few properties of “r” are listed as follows:

1. Range of r: -1 to +1
2. Perfect positive relationship: +1
3. Perfect negative relationship: -1
4. No Linear relationship: 0

PREDICTIVE ANALYTICS

The term "regression" was coined by Francis Galton in 1877 Father of regression Carl F. Gauss (1777-1855).

Definition: A numerical target attribute A collection of data objects also characterized by the target attribute.

The regression task finds a model that allows predicting the target variable value of new objects through $y=f(x_1, x_2, \dots, x_n)$

Regression analysis can be classified based on

Number of explanatory variables

- ♣ Simple regression: single explanatory variable
- ♣ Multiple regression: includes any number of explanatory variables

Types of relationship

- ♣ Linear regression: straight-line relationship
- ♣ Non-linear: implies curved relationships (e.g., logarithmic relationships)

Simple linear regression: $y= \beta_0+ \beta_1 x$

The regression line provides an interpretable model of the phenomenon under analysis

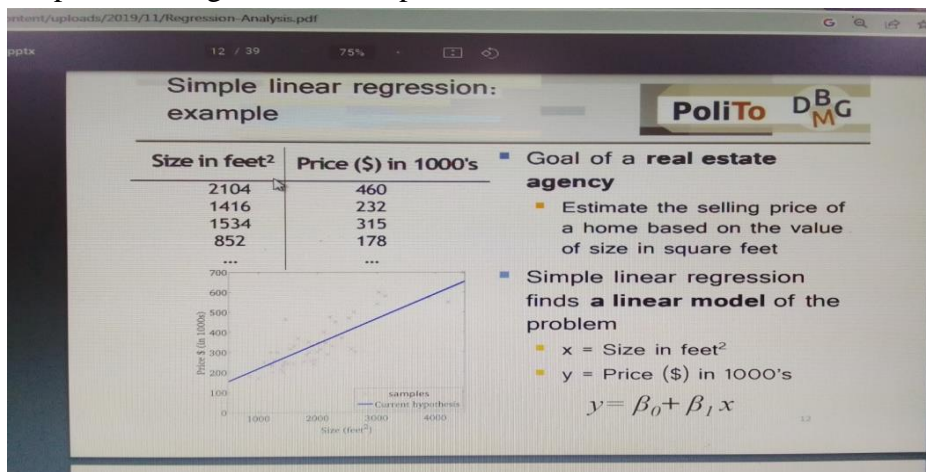
y: estimated (or predicted) value

β_0 : estimation of the regression intercept The intercept represents the estimated value of y when x assumes 0

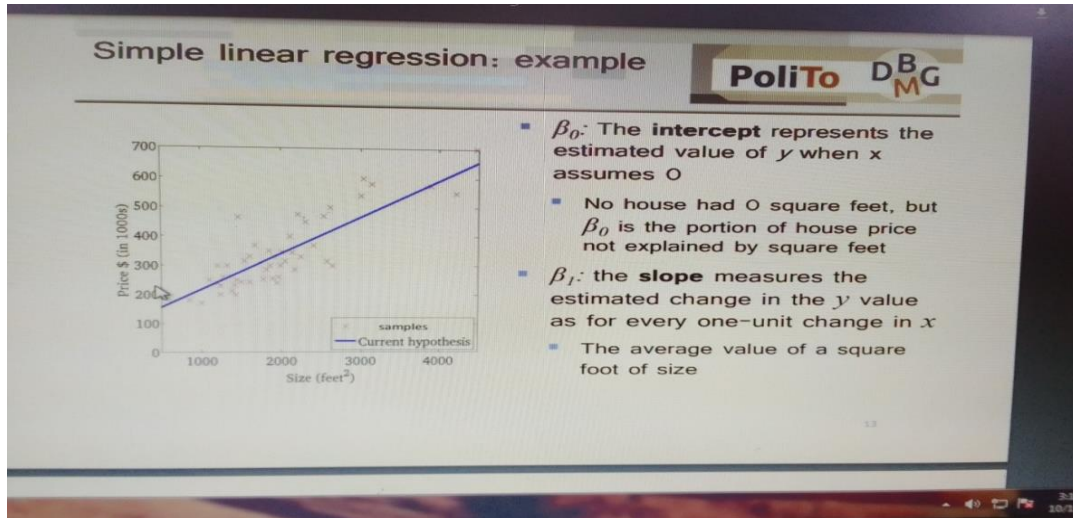
β_1 : estimation of the regression slope

x: independent variable

Simple linear regression: example



PREDICTIVE ANALYTICS



Multiple Regression Definition

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

$$y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

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

Multiple Regression Analysis Definition

Multiple regression analysis permits to control explicitly for many other circumstances that concurrently influence the dependent variable. The objective of regression analysis is to model the relationship between a dependent variable and one or more independent variables. Let k represent the number of variables and denoted by $x_1, x_2, x_3, \dots, x_k$. Such an equation is useful for the prediction of value for y when the values of x are known.

To perform a regression analysis, first calculate the multiple regression of your data. You can use this formula:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

In this formula:

PREDICTIVE ANALYTICS

Y stands for the predictive value or dependent variable.

The variables (X1), (X2) and so on through (Xp) represent the predictive values, or independent variables, causing a change in Y. It's important to note that each X factor represents a distinct predictive value.

The variable (b0) represents the Y-value when all the independent variables (X1 through Xp) are equal to zero.

The variables (b1) through (bp) represent the regression coefficients.

Benefits of multiple regression analysis

- Multiple regression analysis helps us to better study the various predictor variables at hand.
- It increases reliability by avoiding dependency on just one variable and have more than one independent variable to support the event.
- Multiple regression analysis permits you to study more formulated hypotheses that are possible.

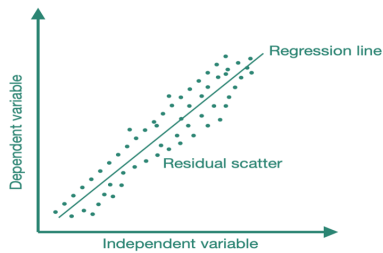
The multiple regression models is based on the following assumptions:

- There is a [linear relationship](#) between the dependent variables and the independent variables
- The independent variables are not too highly [correlated](#) with each other
- y_i observations are selected independently and randomly from the population
- Residuals should be [normally distributed](#) with a mean of 0 and variance
- There should be proper specification of the model in multiple regression. This means that only relevant variables must be included in the model and the model should be reliable.
- Linearity must be assumed; the model should be linear in nature.
- Normality must be assumed in multiple regression. This means that in multiple regression, variables must have normal distribution.
- [Homoscedasticity](#) must be assumed; the variance is constant across all levels of the predicted variable.

Homoscedasticity means “having the same scatter.” For it to exist in a set of data, the points must be about the same distance from the line, as shown in the picture above. The opposite is *heteroscedasticity* (“different scatter”), where points are at widely varying distances from the [regression line](#).

PREDICTIVE ANALYTICS

Homoscedasticity Residual Plot



WallStreetMojo

SLR Model Building: -

1. Collect and extract the data
2. Pre-process the data
3. Divide the data into validation and training data
4. Perform descriptive analytics
5. Define the functional form of regression
6. Estimate the regression parameter
7. Perform regression model diagnostic
8. Validate the model using validation data
9. Decide on the model deployment.

Multiple Linear Regression Model

A multiple linear regression model is a linear equation that has the general form: $y = b_1x_1 + b_2x_2 + \dots + c$ where y is the dependent variable, x_1, x_2, \dots are the independent variable, and c is the (estimated) intercept.

City	Number of weekly riders	Price per week (\$)	Population of city	Monthly income of riders (\$)	Average parking rates per month (\$)
1	192000	15	1800000	5800	50
2	190400	15	1790000	6200	50
3	191200	15	1780000	6400	60
4	177600	25	1778000	6500	60
5	176800	25	1750000	6550	60
6	178400	25	1740000	6580	70
7	180800	25	1725000	8200	75
8	175200	30	1725000	8600	75
9	174400	30	1720000	8800	75
10	173920	30	1705000	9200	80
11	172800	30	1710000	9630	80
12	163200	40	1700000	10570	80
13	161600	40	1695000	11330	85
14	161600	40	1695000	11600	100
15	160800	40	1690000	11800	105
16	159200	40	1630000	11830	105
17	148800	65	1640000	12650	105
18	115696	102	1635000	13000	110
19	147200	75	1630000	13224	125
20	150400	75	1620000	13766	130
21	152000	75	1615000	14010	150
22	136000	80	1605000	14468	155
23	126240	86	1590000	15000	165
24	123888	98	1595000	15200	175
25	126080	87	1590000	15600	175
26	151680	77	1600000	16000	190
27	152800	63	1610000	16200	200

PREDICTIVE ANALYTICS

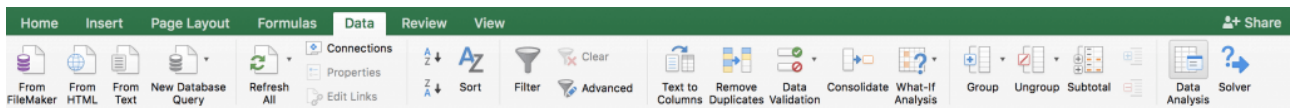
In the above data, the ‘Number of weekly riders’ is a dependent variable that depends on the ‘Price per week (\$)’, ‘Population of city’, ‘Monthly income of riders (\$)’, ‘Average parking rates per month (\$)’.

Let us assign the variables:

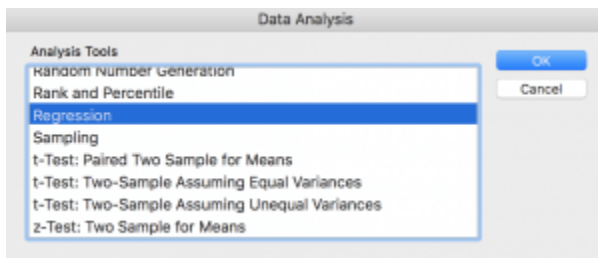
- Price per week (\$) – x_1
- Population of city – x_2
- Monthly income of riders (\$) – x_3
- Average parking rates per month (\$) – x_4
- Number of weekly riders – y

The linear model would be of the form: $y = ax_1 + bx_2 + cx_3 + dx_4 + e$ where a, b, c, d are the respective coefficients and e is the intercept.

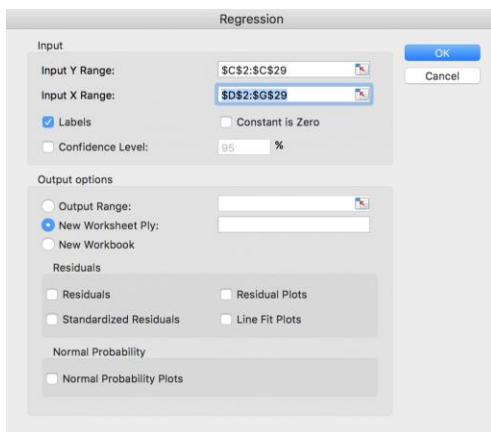
There are two different ways to create the linear model on Microsoft Excel. In this article, we will take a look at the Regression function included in the Data Analysis ToolPak. After the Data Analysis ToolPak has been enabled, you will be able to see it on the Ribbon, under the Data tab:



Click Data Analysis to open the Data Analysis ToolPak, and select Regression from the Analysis tools that are displayed.



Select the data ranges in the options:



PREDICTIVE ANALYTICS

The output looks like this:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.972364343
R Square	0.945492416
Adjusted R Square	0.935581946
Standard Error	5406.370168
Observations	27

ANOVA

	df	SS	MS	F	Significance F
Regression	4	11154120959	2788530240	95.403389	1.43862E-13
Residual	22	643034444.7	29228838.4		
Total	26	11797155404			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	100222.5607	135917.874	0.73737587	0.46868594	-181653.8578	382098.979	-181653.86	382098.979
Price per week (\$)	-689.5227228	95.40286728	-7.2274843	3.0523E-07	-887.3761599	-491.66929	-887.37616	-491.66929
Population of city	0.05494128	0.072339149	0.75949581	0.4556178	-0.095080932	0.20496349	-0.0950809	0.20496349
Monthly income of riders (\$)	-1.301366867	1.627450021	-0.7996355	0.43247166	-4.676491635	2.0737579	-4.6764916	2.0737579
Average parking rates per month (\$)	152.4563673	73.86296237	2.0640435	0.0510037	-0.726041104	305.638776	-0.7260411	305.638776

Right on top are the Regression Statistics. Here we are interested in the following measures:

- **Multiple R**, which is the coefficient of linear correlation
- **Adjusted R Square**, which is the R Square (coefficient of determination) adjusted for more than one independent variable

One Model Building Strategy

The first step

Decide on the type of model that is needed in order to achieve the goals of the study. In general, there are five reasons one might want to build a regression model. They are:

- For **predictive** reasons — that is, the model will be used to predict the response variable from a chosen set of predictors.
- For **theoretical** reasons — that is, the researcher wants to estimate a model based on a known theoretical relationship between the response and predictors.
- For **control** purposes — that is, the model will be used to control a response variable by manipulating the values of the predictor variables.
- For **inferential** reasons — that is, the model will be used to explore the strength of the relationships between the response and the predictors.
- For **data summary** reasons — that is, the model will be used merely as a way to summarize a large set of data by a single equation.

The second step

Decide which predictor variables and response variable on which to collect the data. Collect the data.

The third step

Explore the data. That is:

- On a univariate basis, check for outliers, gross data errors, and missing values.
- Study bivariate relationships to reveal other outliers, to suggest possible transformations, and to identify possible multicollinearities.

I can't possibly over-emphasize the importance of this step. There's not a data analyst out there who hasn't made the mistake of skipping this step and later regretting it when a data point was found in error, thereby nullifying hours of work.

The fourth step

Randomly divide the data into a training set and a validation set:

- The **training set**, with at least 15-20 error degrees of freedom, is used to estimate the model.
- The **validation set** is used for cross-validation of the fitted model.

The fifth step

Using the training set, identify several candidate models:

- Use best subsets regression.
- Use stepwise regression, which of course only yields one model unless different alpha-to-remove and alpha-to-enter values are specified.

The sixth step

Select and evaluate a few "good" models:

- Select the models based on the criteria we learned, as well as the number and nature of the predictors.
- Evaluate the selected models for violation of the model conditions.
- If none of the models provide a satisfactory fit, try something else, such as collecting more data, identifying different predictors, or formulating a different type of model.

The seventh and final step

Select the final model:

- Compare the competing models by cross-validating them against the validation data.
- The model with a smaller mean square prediction error (or larger cross-validation R^2) is a better predictive model.

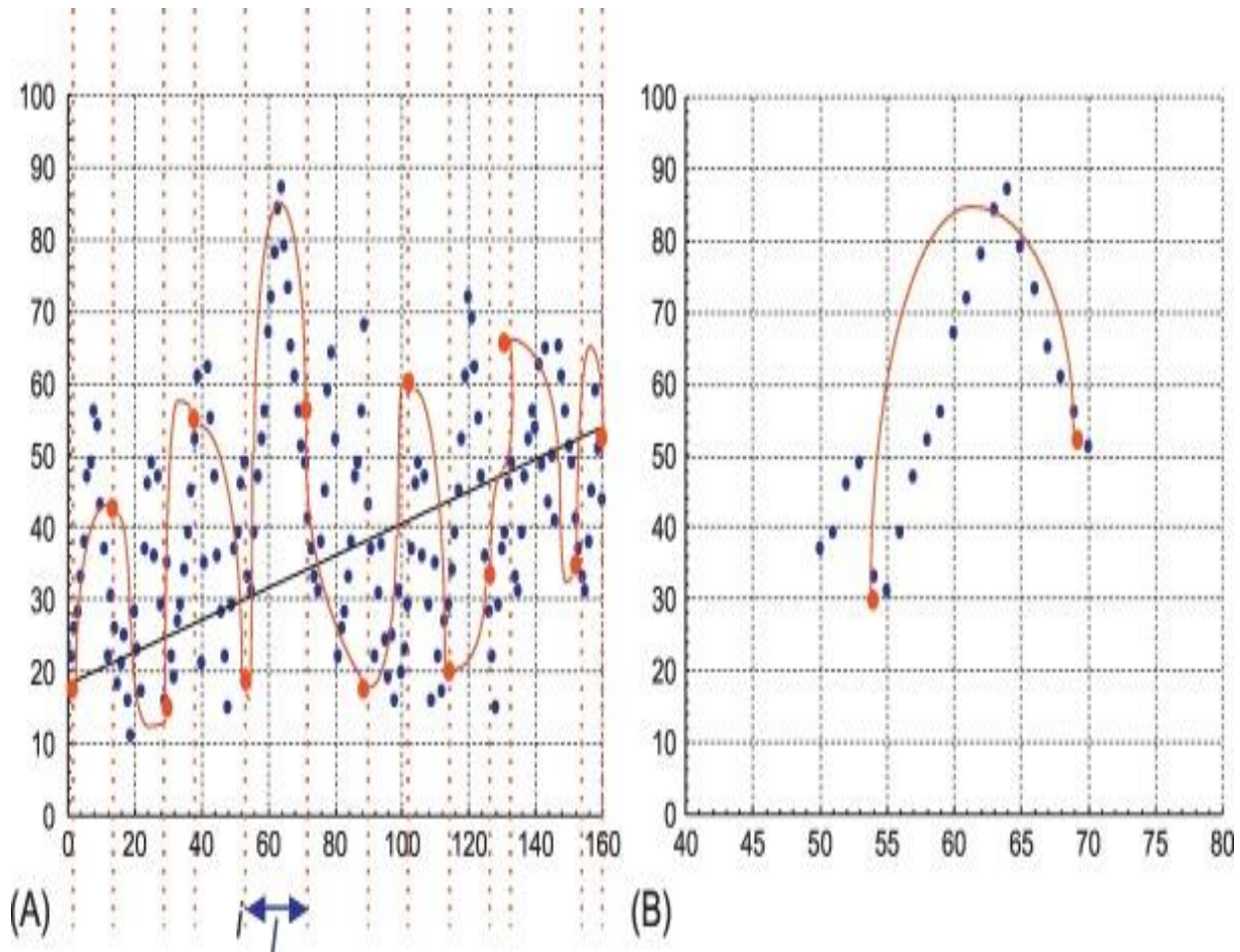
MLR model Building: -

- 1-Collect and extract data
- 2-Pre-process the data
- 3-Perform descriptive analytics
- 4-Decide on the modelling strategy
- 5-Divide the data set into training and validation
- 6-Define the functional form of regression
- 7-Estimate the regression parameters
- 8-Perform regression model diagnostic
- 9-Validate the model using validation data

PREDICTIVE ANALYTICS

UNIT-2

Nonlinear regression refers to a regression analysis where the regression model portrays a nonlinear relationship between a dependent variable and independent variables. In other words, the relationship between predictor and response variable follows a nonlinear pattern.



The simplest statistical relationship between a dependent variable Y and one or more independent or predictor variables X_1, X_2, \dots is

$$Y = B_0 + B_1X_1 + B_2X_2 + \dots + \varepsilon$$

where ε represents a random deviation from the mean relationship represented by the rest of the model. With a single predictor, the model is a straight line. With more than one predictor, the model is a plane or hyperplane. While such models are adequate for representing many relationships (at least over a limited range of the predictors), there are many cases when a more complicated model is required.

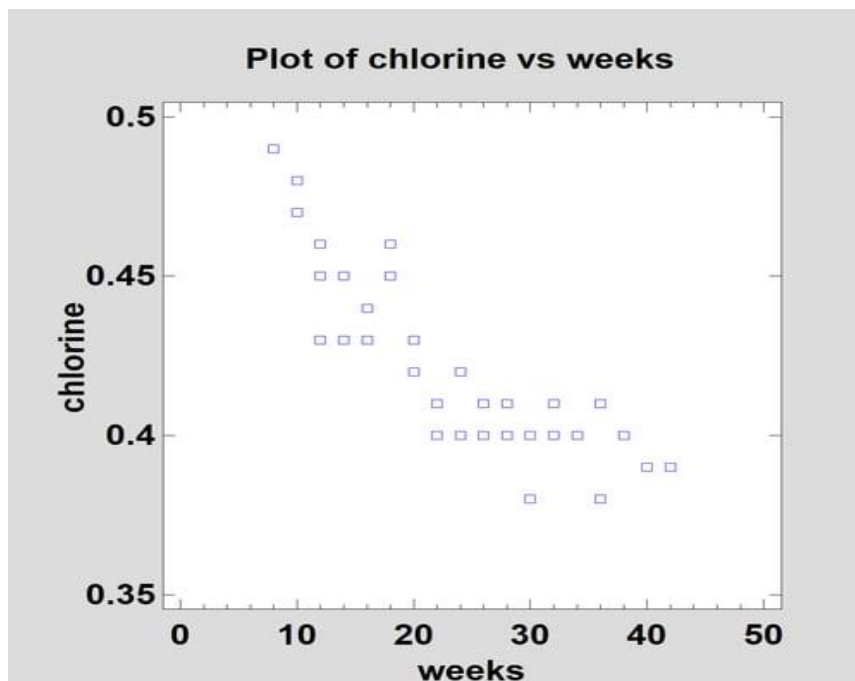
PREDICTIVE ANALYTICS

In Stat graphics, there are several procedures for fitting nonlinear models. The models that may be fit include:

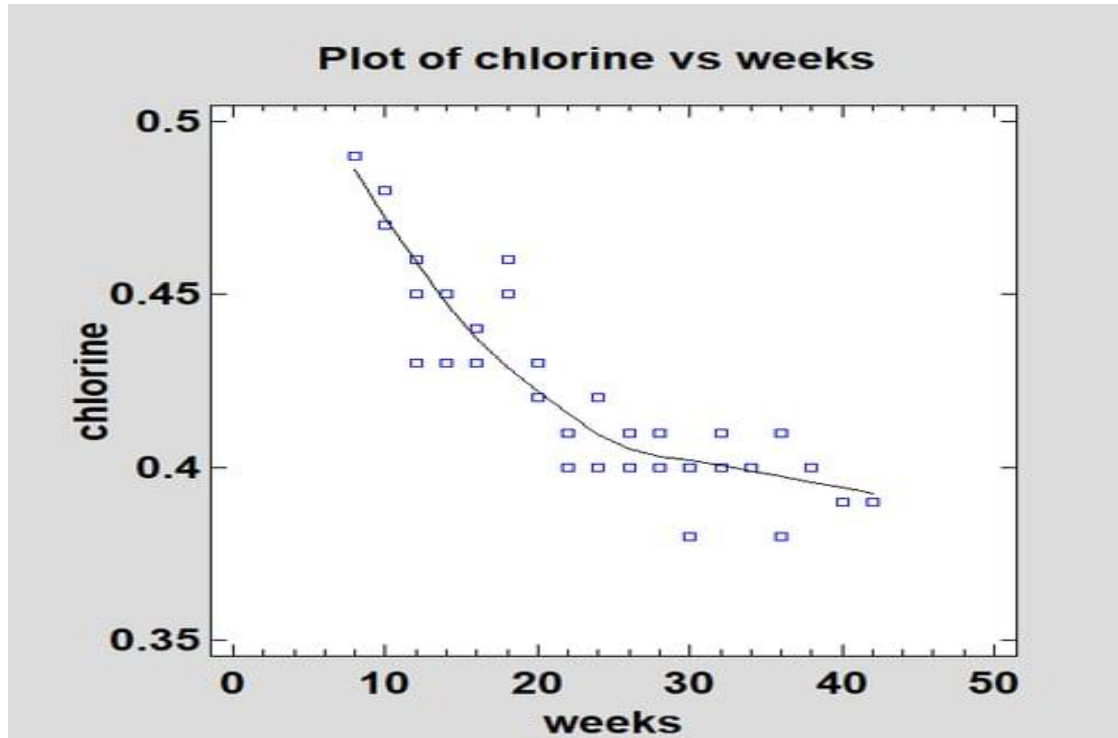
1. **Transformable nonlinear models:** models involving a single predictor variable in which transforming Y, X or both results in a linear relationship between the transformed variables.
2. **Polynomial models:** models involving one or more predictor variables which include higher-order terms such as $B_{1,1}X_1^2$ or $B_{1,2}X_1X_2$.
3. **Models that are nonlinear in the parameters:** models in which the partial derivatives of Y with respect to the predictor variables involve the unknown parameters.

Fitting Transformable Nonlinear Models

In their classic book on regression analysis titled *Applied Regression Analysis*, Draper and Smith show a data set containing 44 samples of a product in which the active ingredient was chlorine. Researchers wanted to model the loss of chlorine as a function of the number of weeks since the sample was produced. As is evident in the scatterplot below, chlorine decays with time.



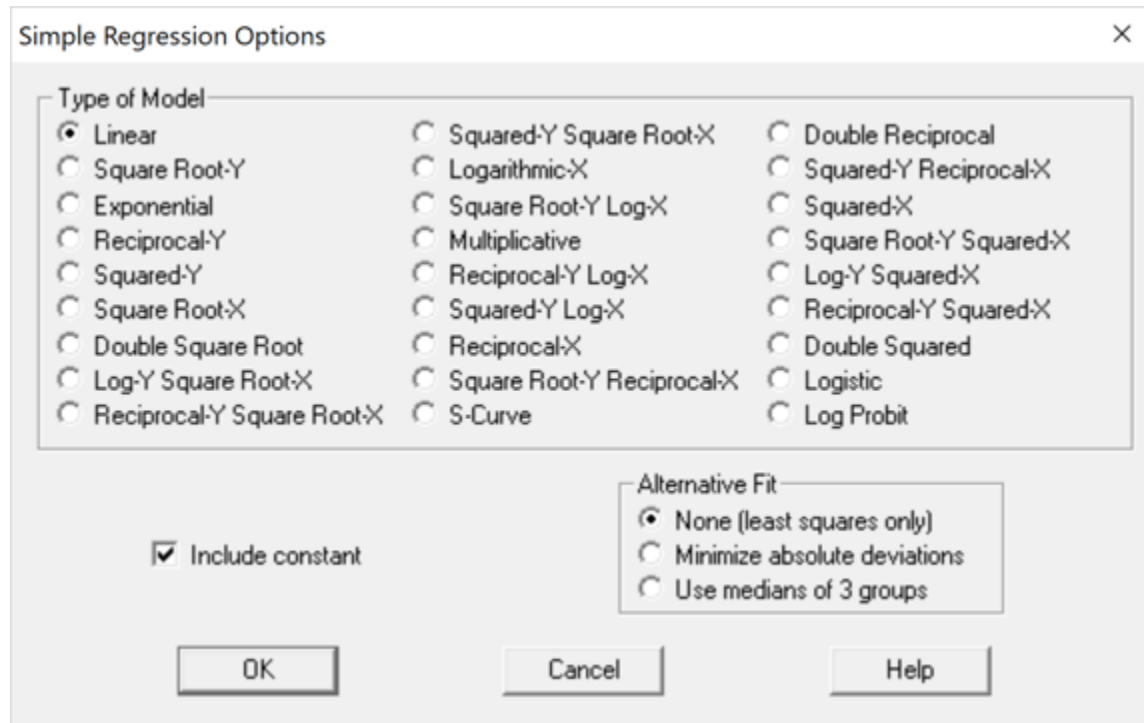
In order to get a quick feel for the shape of the relationship, a robust Lowess smooth may be added to the plot:



Lowess stands for "Locally Weighted Scatterplot Smoothing" and was developed by Bill Cleveland. It smooths the scatter plot by fitting a linear regression at many points along the X axis, weighting observations according to their distance from that point. The procedure is then applied a second time after down-weighting observations that were far removed from the result of the first smooth. It may be seen that there is significant nonlinearity in the relationship between chlorine and weeks.

The *Simple Regression* procedure in Stat graphics gives a choice of many nonlinear functions that may be fit to this data:

PREDICTIVE ANALYTICS



Each function has a form such that after transforming Y, X or both appropriately, the model will be linear in the parameters. For example, the multiplicative model takes the form

$$Y = \alpha X^B$$

which may be linearized by taking logs of both variables:

$$\ln(Y) = \ln(\alpha) + B \ln(X)$$

The one caveat in such an approach is that the error term ε is assumed to be additive **after** the model has been linearized.

To help select a good nonlinear model, Stat graphics will fit all of the models and sort them in decreasing order of R-squared:

PREDICTIVE ANALYTICS

Model	Correlation	R-Squared
Squared-Y reciprocal-X	0.9367	87.75%
Reciprocal-X	0.9333	87.11%
Square root-Y reciprocal-X	0.9312	86.71%
S-curve model	0.9288	86.27%
Double reciprocal	-0.9233	85.25%
Reciprocal-Y logarithmic-X	0.9219	84.99%
Multiplicative	-0.9218	84.98%
Logarithmic-X	-0.9207	84.77%
Squared-Y logarithmic-X	-0.9185	84.36%
Reciprocal-Y square root-X	0.9038	81.69%
Logarithmic-Y square root-X	-0.9012	81.21%
Square root-X	-0.8974	80.54%
Squared-Y square root-X	-0.8926	79.68%
Reciprocal-Y	0.8759	76.73%
Exponential	-0.8710	75.87%
Square root-Y	-0.8682	75.37%
Logistic	-0.8665	75.08%
Log probit	-0.8662	75.03%
Linear	-0.8651	74.83%
Squared-Y	-0.8581	73.63%
Reciprocal-Y squared-X	0.8023	64.37%
Logarithmic-Y squared-X	-0.7941	63.05%
Square root-Y squared-X	-0.7896	62.34%
Squared-X	-0.7849	61.60%
Double squared	-0.7748	60.04%
Double square root	<no fit>	
Square root-Y logarithmic-X	<no fit>	

The R-squared displayed is calculated in the transformed metric, so it represents how well a straight line fits the transformed data. Models near the top of the list are worth considering as alternatives to a linear model. The *Squared-Y reciprocal-X* model has the form

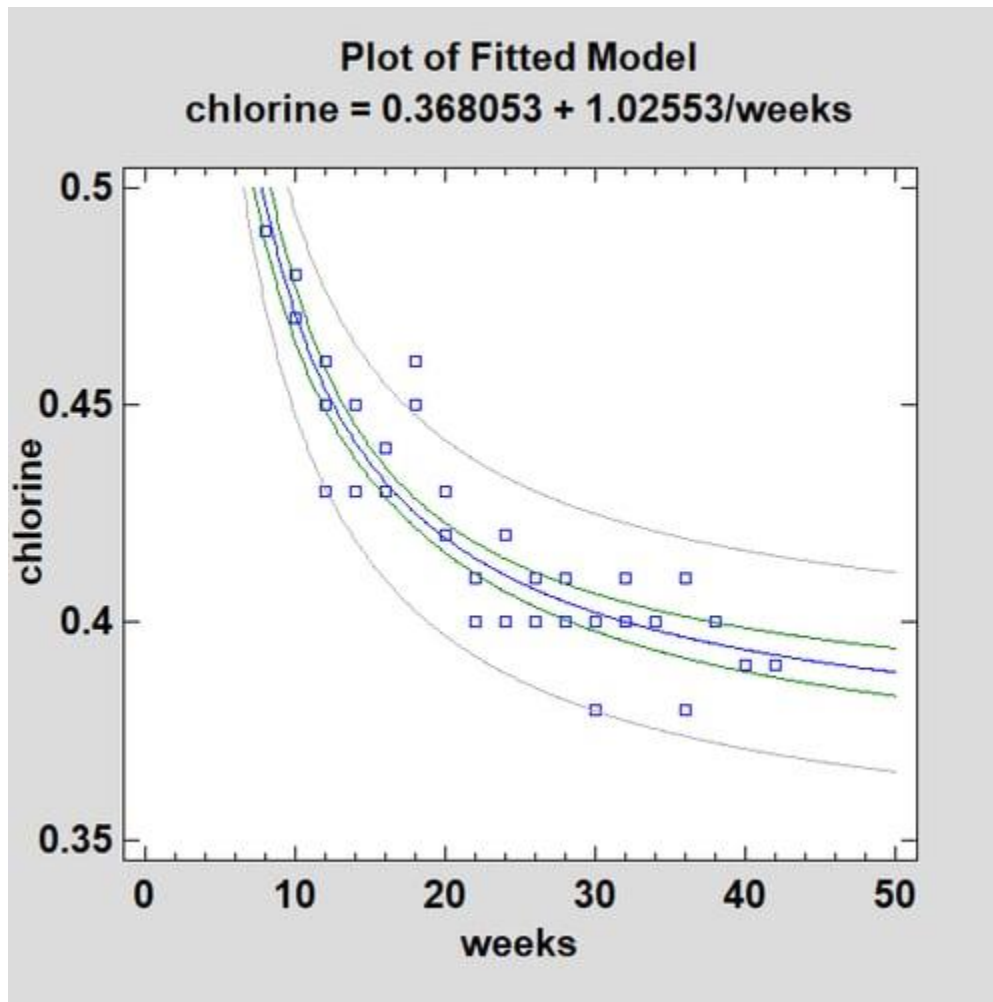
PREDICTIVE ANALYTICS

$$Y^2 = B_0 + B_1/X$$

While the *Reciprocal-X* model is

$$Y = B_0 + B_1/X$$

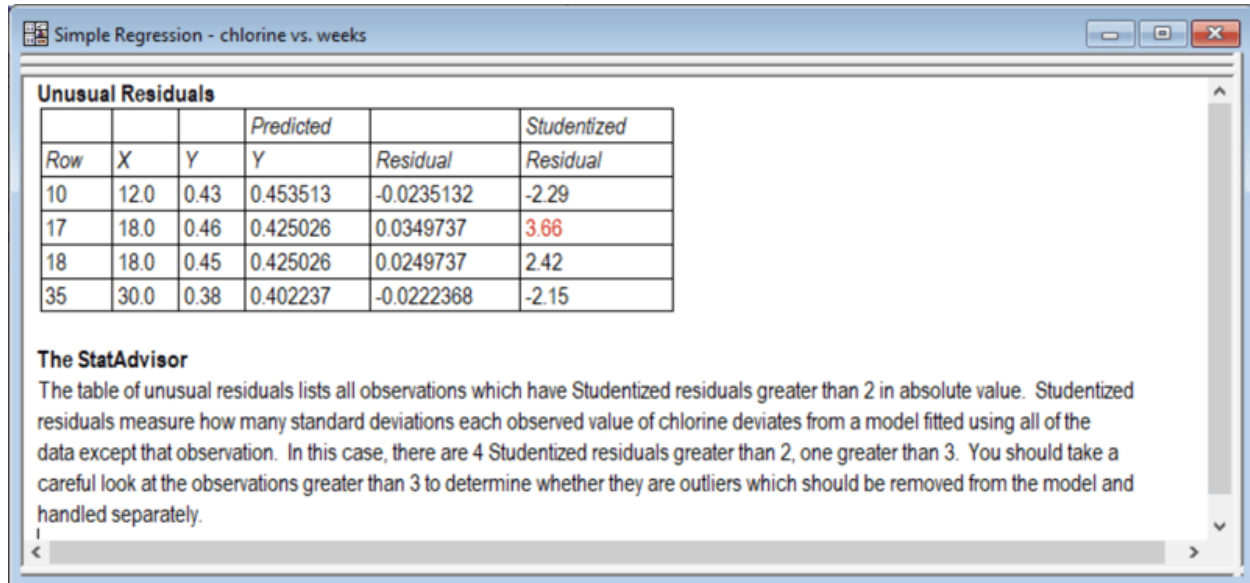
When I'm building empirical models and the results of 2 models are very similar, I usually pick the simpler of the two. Fitting a *Reciprocal-X* model to this data gives the following curve:



In addition to fitting the general relationship well, this model has the pleasing property of reaching an asymptotic value of 0.368053 when *weeks* becomes very large.

Draper and Smith noted the 2 apparent outliers at *weeks* = 18. The Statgraphics *Table of Unusual Residuals* shows that the Studentized residuals for those observations both exceed 2.4:

PREDICTIVE ANALYTICS



The screenshot shows a window titled "Simple Regression - chlorine vs. weeks". It contains a table of "Unusual Residuals" and a "StatAdvisor" section. The table has columns for Row, X, Y, Predicted Y, Residual, and Studentized Residual. Row 17 is highlighted in red, indicating a significant outlier with a Studentized Residual of 3.66. The StatAdvisor text explains that observations with Studentized residuals greater than 2 in absolute value are listed, and notes that there are 4 such observations, one of which is greater than 3.

Row	X	Y	Predicted Y	Residual	Studentized Residual
10	12.0	0.43	0.453513	-0.0235132	-2.29
17	18.0	0.46	0.425026	0.0349737	3.66
18	18.0	0.45	0.425026	0.0249737	2.42
35	30.0	0.38	0.402237	-0.0222368	-2.15

The StatAdvisor
The table of unusual residuals lists all observations which have Studentized residuals greater than 2 in absolute value. Studentized residuals measure how many standard deviations each observed value of chlorine deviates from a model fitted using all of the data except that observation. In this case, there are 4 Studentized residuals greater than 2, one greater than 3. You should take a careful look at the observations greater than 3 to determine whether they are outliers which should be removed from the model and handled separately.

In particular, row #17 is 3.66 standard deviations from its predicted value. However, since they could find no assignable cause that would justify removing those points, Draper and Smith left them in the dataset.

Fitting Polynomial Models

Rather than transforming Y and/or X, we might try fitting a polynomial to the data instead. For example, a second-order polynomial would take the form

$$Y = B_0 + B_1X + B_2X^2$$

while a third-order polynomial would take the form

$$Y = B_0 + B_1X + B_2X^2 + B_3X^3$$

Since polynomials are able to approximate the shape of many curves, they might give a good fit.

The *Polynomial Regression* procedure in Statgraphics fits polynomial models involving a single Y and a single X. The *Analysis Options* dialog box lets the user specify both the order of the polynomial and a shift parameter Δ :

PREDICTIVE ANALYTICS

Polynomial Regression Options

Order: OK

Shift: Cancel

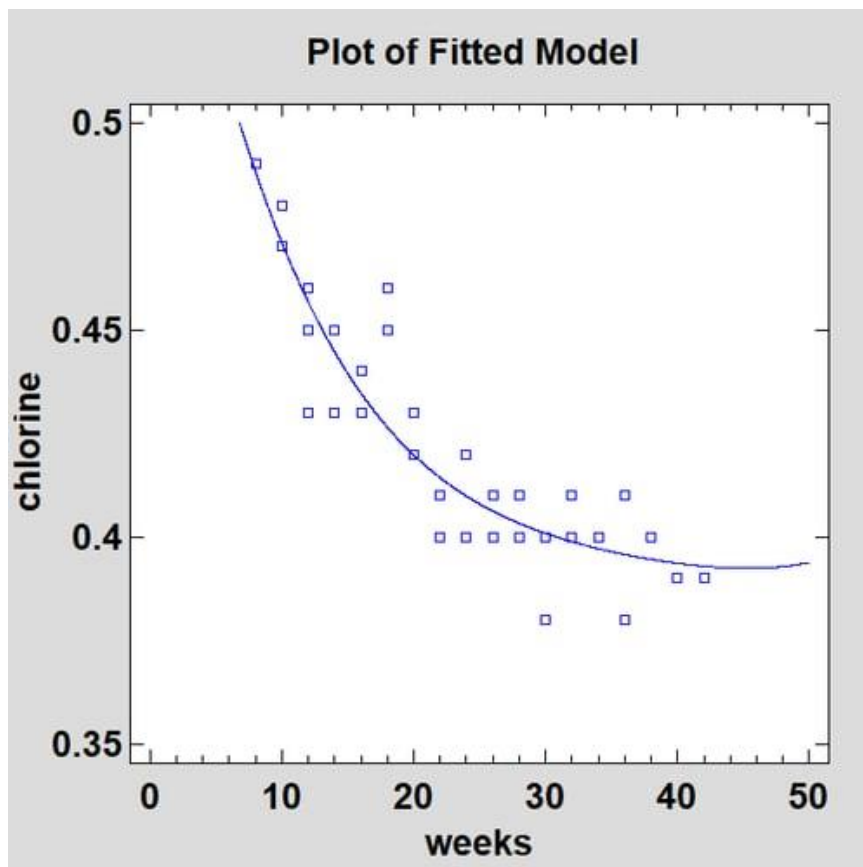
Help

A fourth-order model with a non-zero shift parameter takes the form

$$Y = B_0 + B_1(X-\Delta) + B_2(X-\Delta)^2 + B_3(X-\Delta)^3 + B_4(X-\Delta)^4$$

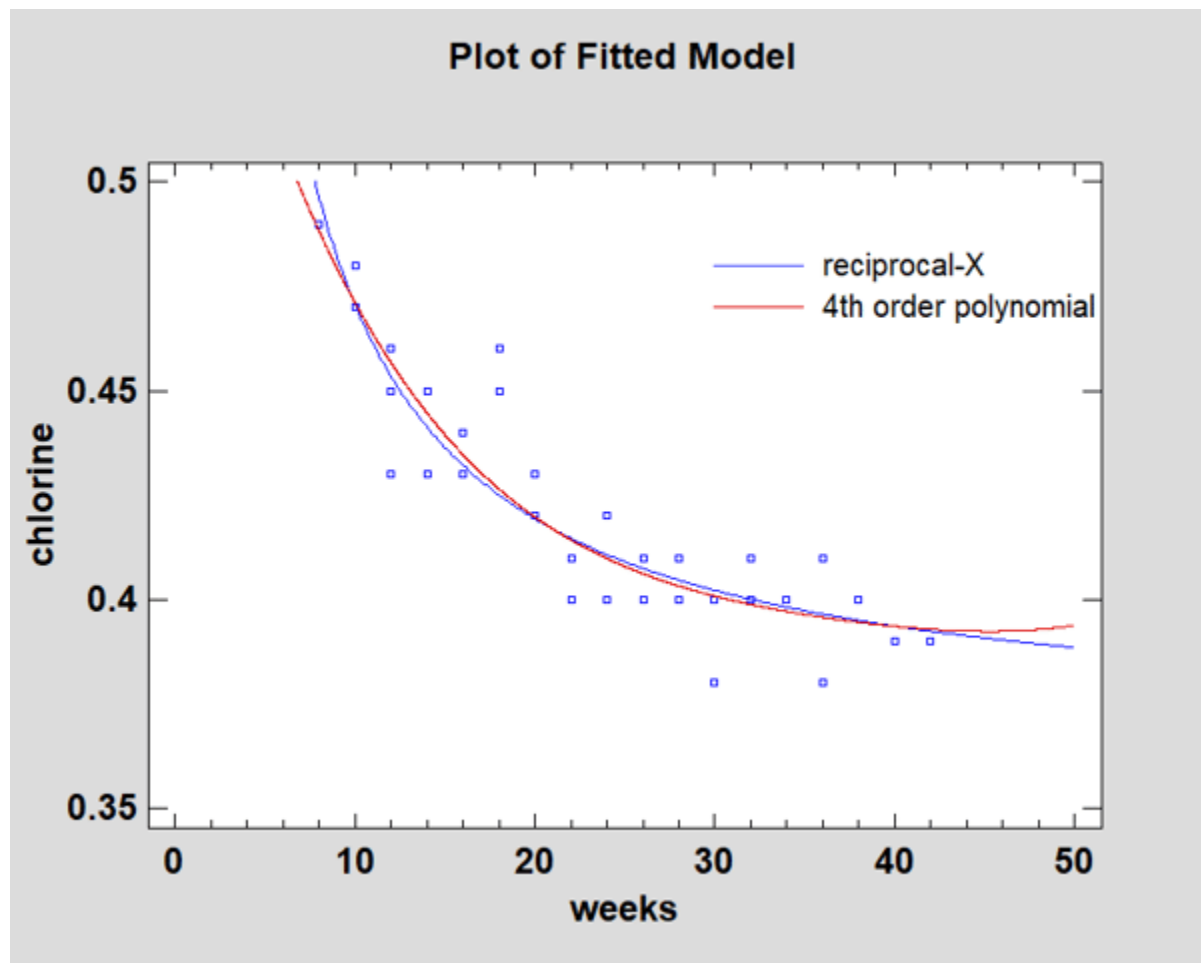
By specifying a non-zero value for Δ , the origin of the polynomial is shifted to a different value of X which can prevent the powers from becoming so large that they overflow the variables created to hold them when performing calculations. Since the maximum value of X is not large in our sample data, the shift parameter may be set equal to 0.

For the chlorine, a fourth-order polynomial fits the data quite well:



PREDICTIVE ANALYTICS

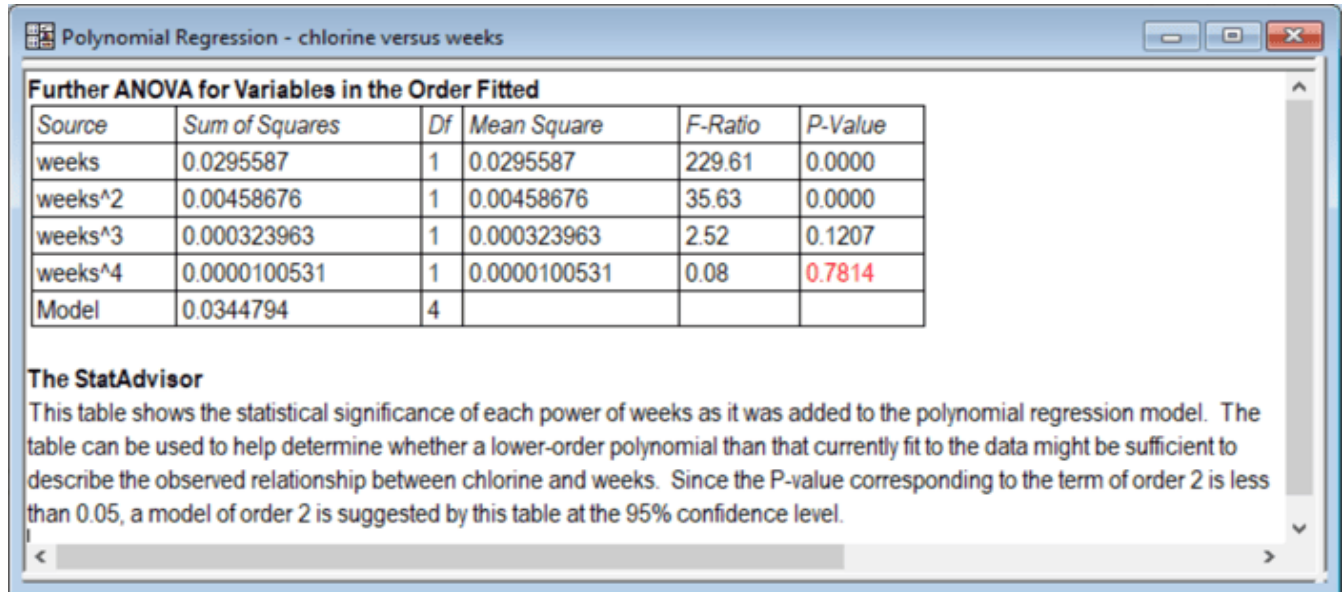
In fact, if we overlay the *Reciprocal-X* model and the fourth-order polynomial in the StatGallery, the predictions are very similar throughout the range of the data:



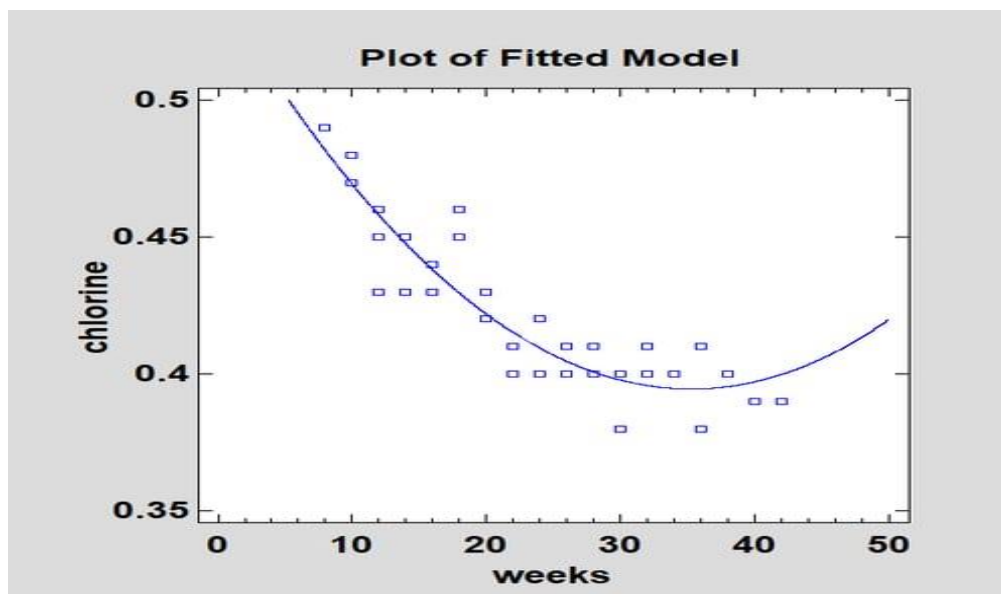
However, beyond the range of the data the polynomial will behave erratically. While the polynomial is suitable if we are only doing interpolation, the *Reciprocal-X* model would be preferred if extrapolation is required.

From a statistical point of view, the 4th order polynomial may be more complicated than is required. Statgraphics creates a table that may be used to help determine what order of polynomial is needed to sufficiently capture the relationship between Y and X. Called the *Conditional Sums of Squares* table, it tests the statistical significance of each term in the polynomial when it is added to a polynomial of one degree less:

PREDICTIVE ANALYTICS



For example, when X^2 is added to a linear model, the P-Value for B^2 equals 0.0000, implying that it significantly improves the fit. When X^3 is added to a second-order model, the P-Value for B^3 equals 0.1207, implying that it does not significantly improve the fit at the 10% significance level. In this case, the P-Values suggest that a second-order polynomial would be sufficient. However, a plot of the fitted model might give one pause:



Even if only using the model for interpolation, the curvature in the interval between 30 and 40 weeks is disconcerting.

PREDICTIVE ANALYTICS

LINEAR REGRESSION EQUATIONS

Linear regression requires a linear model. But what does that really mean?

A model is linear when each term is either a [constant](#) or the product of a [parameter](#) and a [predictor variable](#). A linear equation is constructed by adding the results for each term. This constrains the equation to just one basic form:

Response = constant + parameter * predictor + ... + parameter * predictor

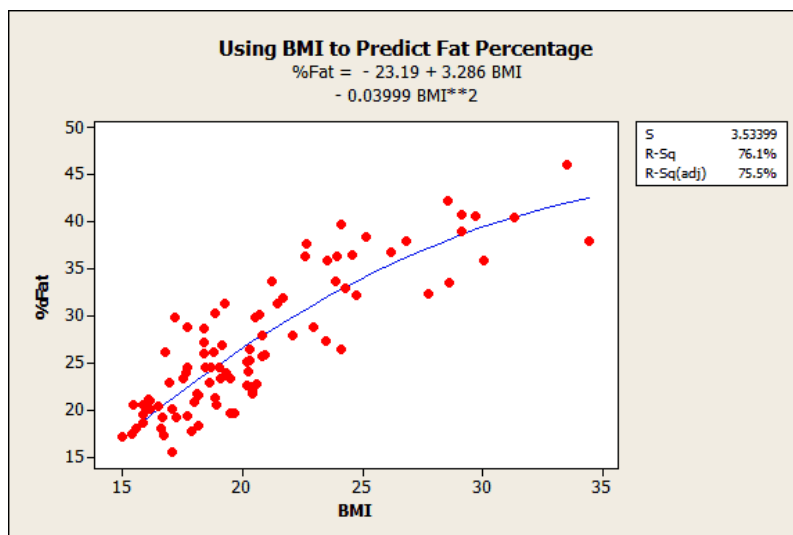
$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

In statistics, a regression equation (or function) is linear when it is linear in the parameters. While the equation must be linear in the parameters, you can transform the predictor variables in ways that produce curvature. For instance, you can include a squared variable to produce a U-shaped curve.

$$Y = b_0 + b_1X_1 + b_2X_1^2$$

This model is still linear in the parameters *even though the predictor variable is squared*. You can also use log and inverse functional forms that are linear in the parameters to produce different types of curves.

Here is an example of a linear regression model that uses a squared term to fit the [curved relationship between BMI and body fat percentage](#).



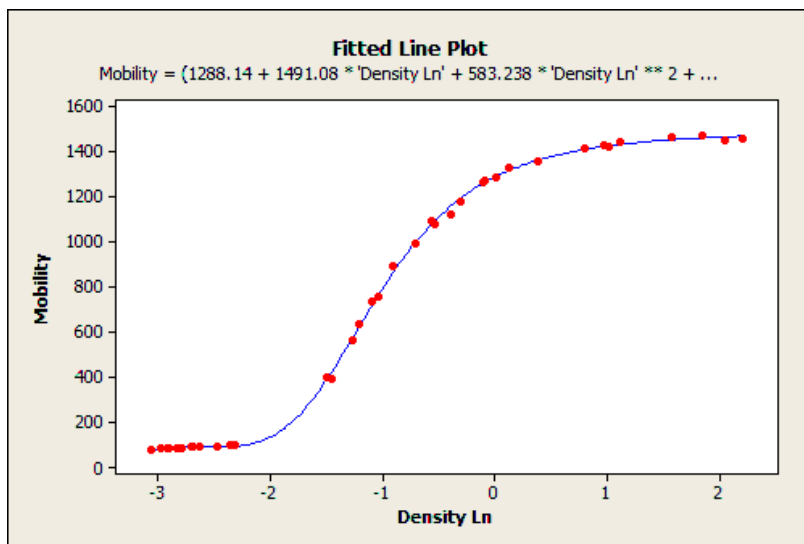
NONLINEAR REGRESSION EQUATIONS

While a linear equation has one basic form, nonlinear equations can take many different forms. The easiest way to determine whether an equation is nonlinear is to focus on the term “nonlinear” itself. Literally, it’s not linear. If the equation doesn’t meet the criteria above for a linear equation, it’s nonlinear.

That covers many different forms, which is why nonlinear regression provides the most flexible curve-fitting functionality. Here are several examples from [Minitab’s](#) nonlinear function catalog. Thetas represent the parameters and X represents the predictor in the nonlinear functions. Unlike linear regression, these functions can have more than one parameter per predictor variable.

Nonlinear function	One possible shape
Power (convex): $\text{Theta1} * X^{\text{Theta2}}$	
Weibull growth: $\text{Theta1} + (\text{Theta2} - \text{Theta1}) * \exp(-\text{Theta3} * X^{\text{Theta4}})$	
Fourier: $\text{Theta1} * \cos(X + \text{Theta4}) + (\text{Theta2} * \cos(2*X + \text{Theta4}) + \text{Theta3})$	

Here is an example of a nonlinear regression model of the [relationship between density and electron mobility](#).



PREDICTIVE ANALYTICS

The nonlinear equation is so long it that it doesn't fit on the graph:

$$\text{Mobility} = (1288.14 + 1491.08 * \text{Density Ln} + 583.238 * \text{Density Ln}^2 + 75.4167 * \text{Density Ln}^3) / (1 + 0.966295 * \text{Density Ln} + 0.397973 * \text{Density Ln}^2 + 0.0497273 * \text{Density Ln}^3)$$

Linear and nonlinear regression are actually named after the functional form of the models that each analysis accepts. I hope the distinction between linear and nonlinear equations is clearer and that you understand how it's possible for linear regression to model curves! It also explains why you'll see R-squared displayed for some curvilinear models even though [it's impossible to calculate R-squared for nonlinear regression](#).

Difference between linear and non linear regression models

LINEAR MODELS	NONLINEAR MODELS
<ul style="list-style-type: none">• Simpler models• All materials comply with Hooke's Law• Few input parameters• Assessment of the global structural behaviour• Faster calculation• Convergence is not a problem in dynamic analyses	<ul style="list-style-type: none">• Complex models• Many input parameters• Assessment of the structural elements behaviour• Slower calculation• Possible convergence problems in dynamic analyses• Needed if the loading produces a significant changes in the stiffness• Important if large displacements change the geometry

Linear Regression Equations

A linear regression model follows a very particular form. In statistics, a regression model is linear when all terms in the model are one of the following:

- The constant
- A [parameter](#) multiplied by an independent variable (IV)

Then, you build the equation by only adding the terms together. These rules limit the form to just one type:

$$\text{Dependent variable} = \text{constant} + \text{parameter} * \text{IV} + \dots + \text{parameter} * \text{IV}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K$$

This type of regression equation is linear in the parameters. However, it is possible to model curvature with this type of model. While the function must be linear in the parameters, you can raise an independent variable by an exponent to fit a curve. For example, if you square an independent variable, the model can follow a U-shaped curve.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2$$

PREDICTIVE ANALYTICS

The regression example below models the relationship between body mass index (BMI) and body fat percent. In a different blog post, I use this model to show [how to make predictions with regression analysis](#). It is a linear model that uses a quadratic (squared) term to model the curved relationship.

MULTICOLLINEARITY:

Collinearity (and Multicollinearity) means that the predictors variables, also known as independent variables, aren't so independent.

Word Anatomy of Multicollinearity

Multi-col-linear-ity

Referring to the multiple independent variables within multiple regression.

A modification of the prefix co, meaning together or joint. Referencing the linear movement in tandem i.e., correlation.

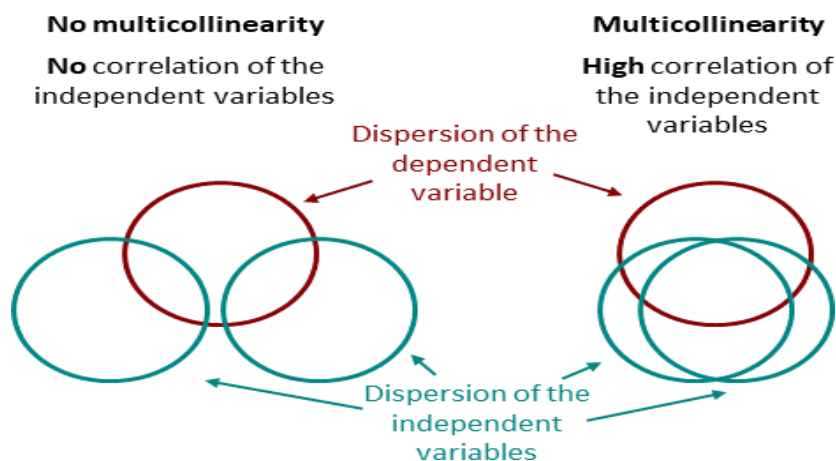
Occurring within a linear equation.

Suffix meaning the quality or state of.

Collinearity is a situation where two features are linearly associated (high correlated), and they are used as predictors for the target. It's often measured using Pearson's correlation coefficient. Collinearity between more than two predictors is also possible (and often the case).

The term multicollinearity was first used by Ragnar Frisch. Multicollinearity is a special case of collinearity where a feature exhibits a linear relationship with two or more features. We can also have a situation where more than two features are correlated and, at the same time, have no high correlation pairwise.

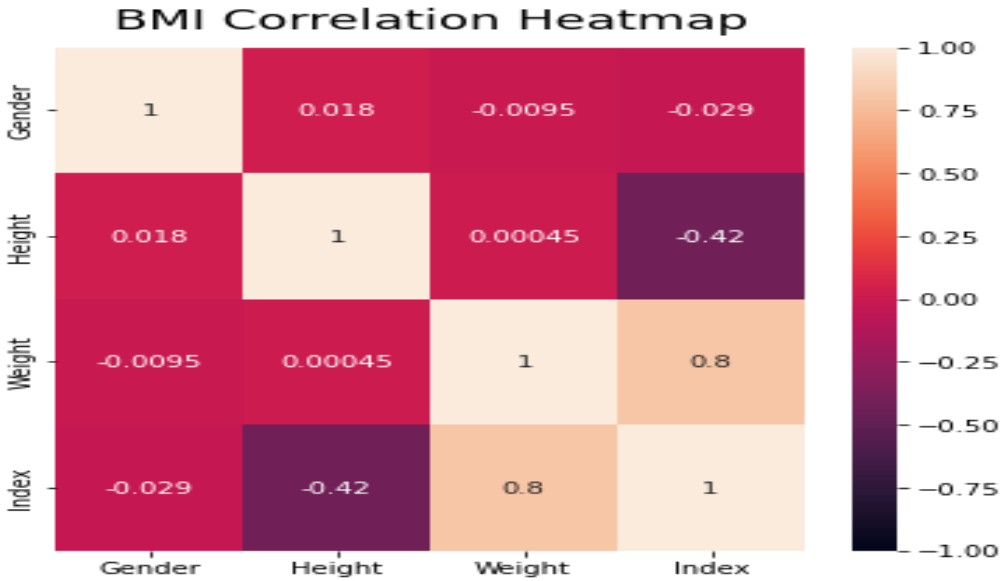
Partial multicollinearity is ubiquitous in multiple regression. Two random variables will almost always correlate at some level in a sample, even if they share no fundamental relationship in the larger population. In other words, multicollinearity is a matter of degree.



PREDICTIVE ANALYTICS

The correlation matrix

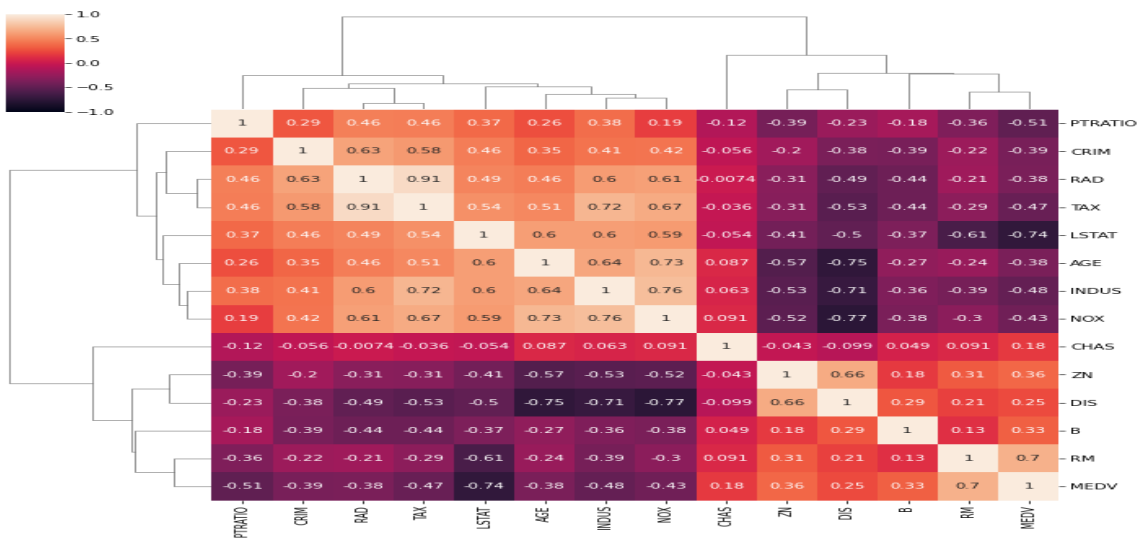
The correlation matrix gives you the pairwise correlation or bivariate relationship between two independent variables - **collinearity**.



We can see a strong correlation between 'Index' and 'Height' / 'Weight' (as expected). We can also notice that 'Weight' has much more impact on INdex than 'Height'. That should be also intuitive and expected.

Clustermap

Clustermap table shows not only all correlation between variables, but also group (cluster) relationships.



PREDICTIVE ANALYTICS

Variance inflation factor

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. VIF is used to identify the correlation of one independent variable with a group of other variables.

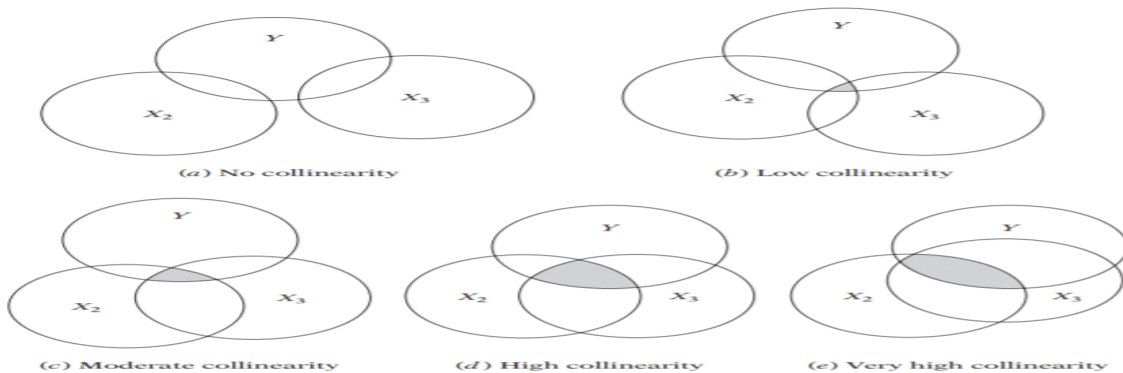
$$VIF_i = 1 / (1 - R_i^2)$$

So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

- VIF starts at 1 and has no upper limit
 - $VIF = 1$, no correlation between the independent variable and the other variables
 - VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others
- | | |
|------------|-----------|
| • Feature | VIF |
| • 0 Gender | 2.028864 |
| • 1 Height | 11.623103 |
| • 2 Weight | 10.688377 |

'Height' and 'Weight' have high values of VIF, indicating that these two variables are highly correlated. This is expected as the height of a person does influence their weight. Hence, considering these two features together still leads to a model with high multicollinearity.

In regression analysis, multicollinearity exists when two or more of the variables demonstrate a linear relationship between them. The VIF measures by how much the **linear correlation** of a given regressor with the other regressors increases the variance of its coefficient estimate with respect to the baseline case of no correlation.



PREDICTIVE ANALYTICS

Two kinds of multicollinearity

1. **Data-based multicollinearity:** this type of multicollinearity is present in the data itself. Observational experiments are more likely to exhibit this kind of multicollinearity.
 - Example: two identical (or almost identical) variables. Weight in pounds and weight in kilos, or investment income and savings/bond income.
1. **Structural multicollinearity:** caused by the researcher creating new predictor variables. This type occurs when we create a model term using other terms. In other words, it's a byproduct of the model that we specify.
 - Including a variable in the regression that is actually a combination of two other variables. For example, including "total investment income" when total investment income = income from stocks and bonds + income from savings interest.

In regression analysis, multicollinearity has the following types:

1. None: When the regression exploratory variables have no relationship with each other, then there is no multicollinearity in the data.
2. Low: When there is a relationship among the exploratory variables, but it is very low, then it is a type of low multicollinearity.
3. Moderate: When the relationship among the exploratory variables is moderate, then it is said to be moderate multicollinearity.
4. High: When the relationship among the exploratory variables is high or there is perfect correlation among them, then it said to be high multicollinearity.
5. Very high: When the relationship among the exploratory variables is exact, then it is the problem of very high multicollinearity, which should be removed from the data when regression analysis is conducted.

AUTOCORRELATION: The autocorrelation function is a statistical representation that can analyze the degree of similarity between a time series and a lagged version of itself. This function allows the analyst to compare the current value of a data set to its past value. To use this function, the analyst uses the same time series and compares it against a lagged version of itself over one or more time periods. They assess the strength of the correlation between these different versions to uncover trends and patterns that allow them to evaluate the strength of the relationship between two or more variables.



PREDICTIVE ANALYTICS

The autocorrelation function often works by producing a value between -1 and +1. A positive autocorrelation represents a positive correlation, which means that as one time series increases, you may see a proportionate increase in the other time series. A negative correlation can mean that an increase in one time series results in a proportionate decrease in the other. The closer to plus or minus one the value falls, the stronger the correlation in either direction.

For example, a weather scientist might use this function when analyzing the minimum daily temperature recorded in a city using a data set from the last 10 years. They insert the data into a statistical modeling program and perform an autocorrelation analysis. The program produces a graph that shows how the minimum daily temperature in the city has changed over the last 10 years, indicating an increase in the daily minimum temperature with a high degree of confidence. This degree of confidence shows that the positive correlation between minimum temperature and time is likely not the result of random chance.

When can you use the autocorrelation function: The autocorrelation function has various uses in many industries that rely on time-related statistical models. Here are a few industries that use autocorrelation functions, with examples of how industry professionals may use these functions to complete their work:

Physics and engineering

Autocorrelation functions have various applications in physics and engineering. In particular, these functions help scientists measure and understand patterns in the behaviors of sound waves and light. For example, a physicist might use this type of function when studying patterns in how light scatters when moving through a particular medium, like air or a liquid. They may also use this function to study sonic concepts like pitch, frequency and tempo. An astrophysicist may use autocorrelation functions to understand how wavelengths travel through space, with consideration for how physical principles like gravity affect their behavior.

Meteorology

Meteorologists and climate researchers frequently use autocorrelation functions. They use this function to understand how weather patterns change over time and how different variables influence these trends. For example, meteorologists use historical data patterns to predict changes in future weather conditions. These scientists create statistical models using autocorrelation functions to assess how weather trends like precipitation and temperature and natural phenomenon like hurricanes have changed over time and how they may continue to change in the future. Understanding weather patterns is important for predicting emergency weather conditions and natural disasters so people can prepare ahead for these events.

Finance

Autocorrelation functions also apply to financial modeling. Stock analysts often use autocorrelation functions to assess trends in a stock's value over time and use that data to predict its future value. Another application of autocorrelation functions in finance is its use in technical analysis. A technical analyst can use autocorrelation functions to understand how past prices for a security may influence its future value. For example, if an autocorrelation function reveals that a stock has accumulated significant gains over two or more days, it's reasonable to predict that the stock may continue to gain in the following days.

PREDICTIVE ANALYTICS

Health and medicine

Autocorrelation functions also have applications in medical technology and research. In particular, many types of medical imaging software depend on algorithms that use autocorrelation functions to operate. For example, ultrasound imaging techniques may use autocorrelation functions to produce visual representations of blood flow in a patient. Another application is in epidemiology tracking and forecasting. Epidemiologists can use autocorrelation functions to identify trends in disease outbreaks within particular regions over a period of time. This can help them understand patterns of outbreaks and devise solutions to minimize or eliminate their impact on vulnerable communities.

How to use autocorrelation functions:

Although the procedure for using an autocorrelation function can depend on the industry and goal, here are some general steps to help you apply these functions to your needs:

1. Determine the time series for analysis

The first step to using the autocorrelation function is to determine your variables and gather your data set for analysis. Your methods for collecting your time series depend on your industry and what kind of data you wish to measure. For example, many meteorologists can access data sets from public sources for information regarding general weather conditions, like temperature and precipitation levels. Those performing more specialized types of research may need to devise their own data collection strategy that suits their specific research needs. For example, biomedical engineers typically develop individualized studies to collect data relevant to their research.

2. Choose a statistical modeling program

Since analysts typically work with large data sets when performing autocorrelation functions, it's common to use a statistical modeling program to store and process the data. Professionals in different industries may use specialized statistical modeling programs for their field. For example, technical analysts working in finance may use a different program than those working in medical research. Using software that's designed to process data in your industry makes it easier to access the analytical tools you need and accurately visualize the results. When choosing modeling software, research ones related to your industry or specialty.

3. Input the data and run the autocorrelation function

After choosing the right modeling software for your data set, upload or input your data into the system. Depending on your program and how extensive your data set is, you may either manually input your data points or upload them all at once. After inputting the data, determine what function to use to calculate autocorrelation for your time series. There are many different formulas for calculating autocorrelation, and choosing the right one to adapt to your specific analysis depends on the purpose of the analysis and how you plan to use the results.

One common type of autocorrelation function is the Durbin-Watson test. This statistic uses regression analysis to identify autocorrelation in a time series. When you apply it, the Durbin-Watson test assesses the degree of correlation between variables in a time series in a range of zero to four. Results closer to zero indicate a stronger positive correlation between the variables, while values closer to four show a stronger negative pattern of correlation. If the value falls

PREDICTIVE ANALYTICS

between zero and four, it suggests less autocorrelation. Although the Durbin-Watson test is common in financial analysis, it may be less common in other industries.

4. Produce a visual representation and interpret results

Once you've applied the function to your time series, most statistical modeling software produces a visual representation, like a graph, to help you interpret the results. Graphs visualize the linearity of the relationships between the variables, making it easy to interpret the degree of correlation by reviewing how the points map onto the graph. For example, the graph may show a strong upward trend in the data points, indicating a clear positive correlation between the variables. If the data points distribute more randomly within the graph, it suggests that there's less of a correlation between the points.

Durbin Watson Statistic

The Durbin Watson (DW) statistic is a test for [autocorrelation](#) in the residuals from a statistical model or [regression analysis](#). The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation and values from 2 to 4 means negative autocorrelation.

- The Durbin Watson statistic is a test for autocorrelation in a regression model's output.
- The DW statistic ranges from zero to four, with a value of 2.0 indicating zero autocorrelation.
- Values below 2.0 mean there is positive autocorrelation and above 2.0 indicates negative autocorrelation.

The Durbin–Watson statistic, while displayed by many regression analysis programs, is not applicable in certain situations.

Example of the Durbin Watson Statistic

The formula for the Durbin Watson statistic is rather complex but involves the residuals from an ordinary [least squares \(OLS\) regression](#) on a set of data. The following example illustrates how to calculate this statistic.

Assume the following (x,y) data points:

Pair One=(10,1,100)

Pair Two=(20,1,200)

Pair Three=(35,985)

Pair Four=(40,750)

Pair Five=(50,1,215)

Pair Six=(45,1,000)

PREDICTIVE ANALYTICS

Using the methods of a least squares regression to find the "[line of best fit](#)," the equation for the best fit line of this data is:

$$Y = -2.6268x + 1,129.2$$

This first step in calculating the Durbin Watson statistic is to calculate the expected "y" values using the line of best fit equation. For this data set, the expected "y" values are:

$$\text{Expected } Y(1) = (-2.6268 \times 10) + 1,129.2 = 1,102.9$$

$$\text{Expected } Y(2) = (-2.6268 \times 20) + 1,129.2 = 1,076.7$$

$$\text{Expected } Y(3) = (-2.6268 \times 35) + 1,129.2 = 1,037.3$$

$$\text{Expected } Y(4) = (-2.6268 \times 40) + 1,129.2 = 1,024.1$$

$$\text{Expected } Y(5) = (-2.6268 \times 50) + 1,129.2 = 997.9$$

$$\text{Expected } Y(6) = (-2.6268 \times 45) + 1,129.2 = 1,011$$

Next, the differences of the actual "y" values versus the expected "y" values, the errors, are calculated:

$$\text{Error}(1) = (1,100 - 1,102.9) = -2.9$$

$$\text{Error}(2) = (1,200 - 1,076.7) = 123.3$$

$$\text{Error}(3) = (985 - 1,037.3) = -52.3$$

$$\text{Error}(4) = (750 - 1,024.1) = -274.1$$

$$\text{Error}(5) = (1,215 - 997.9) = 217.1$$

$$\text{Error}(6) = (1,000 - 1,011) = -11$$

Next these errors must be [squared and summed](#):

$$\text{Sum of Errors Squared} = (-2.92 + 123.32 + -52.32 + -274.12 + 217.12 + -112) = 140,330.81$$

Next, the value of the error minus the previous error are calculated and squared:

$$\text{Difference}(1) = (123.3 - (-2.9)) = 126.2$$

$$\text{Difference}(2) = (-52.3 - 123.3) = -175.6$$

$$\text{Difference}(3) = (-274.1 - (-52.3)) = -221.9$$

PREDICTIVE ANALYTICS

$$\text{Difference}(4)=(217.1-(-274.1))=491.3$$

$$\text{Difference}(5)=(-11-217.1)=-228.1$$

$$\text{Sum of Differences Square}=389,406.71$$

Finally, the Durbin Watson statistic is the quotient of the squared values:

$$\text{Durbin Watson}=389,406.71/140,330.81=2.77$$

Serial Correlation (Autocorrelation)

Serial correlation, also known as autocorrelation, occurs when the regression residuals are correlated with each other. In other words, it occurs when the errors in the regression are not independent of each other. This can happen for various reasons, including incorrect model specification, not randomly distributed data, and misspecification of the error term.

This is common with time-series data. One example of serial correlation is found in stock prices. Stock prices tend to go up and down together over time, which is said to be “serially correlated.” This means that if stock prices go up today, they will also go up tomorrow. Similarly, if stock prices go down today, they are likely to go down tomorrow. The degree of serial correlation can be measured using the autocorrelation coefficient. The **autocorrelation coefficient** measures how closely related a series of data points are to each other.

Types of Serial Correlation

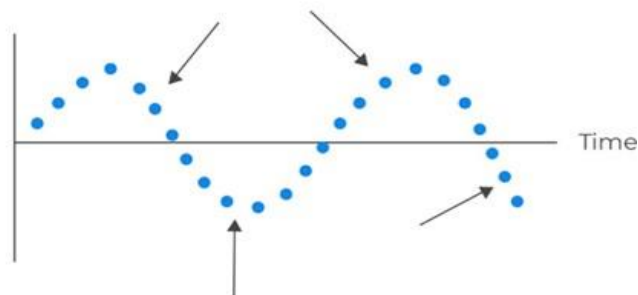
Positive Serial Correlation

Positive serial correlation occurs when a positive error for one observation increases the chance of a positive error for another observation. In other words, if there is a positive error in one period, there is a greater likelihood of a positive error in the next period as well. Positive serial correlation also means that a negative error for one observation increases the chance of a negative error for another observation. So, if there is a negative error in one period, there is a greater likelihood of a negative error in the next period.



Positive Autocorrelation

Above-average errors tend to follow above-average errors



Below-average errors tend to follow below-average errors

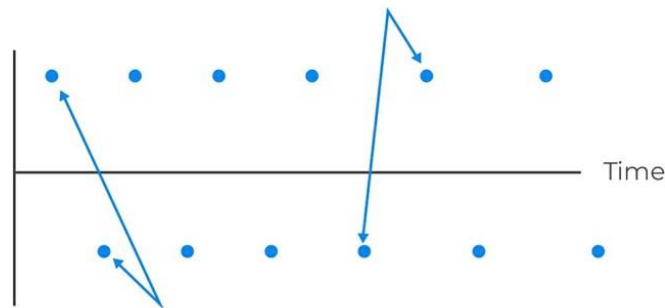
Negative Serial Correlation

A negative serial correlation occurs when a positive error for one observation increases the chance of a negative error for another observation. In other words, if there is a positive error in one period, there is a greater likelihood of a negative error in the next period. A negative serial correlation also means that a negative error for one observation increases the chance of a positive error for another observation. So, if there is a negative error in one period, there is a greater likelihood of a positive error in the next period.



Negative Autocorrelation

Above-average errors tend to follow below-average errors



Below-average errors tend to follow above-average errors

Durbin-Watson Test

The Durbin-Watson test is a statistical test used to determine whether or not there is a serial correlation in a data set. It tests the null hypothesis of no serial correlation against the alternative positive or negative serial correlation hypothesis. The test is named after James Durbin and Geoffrey Watson, who developed it in 1950.

The Durbin-Watson Statistic (DW) is approximated by:

$$DW=2(1-r)$$

Where:

r is the sample correlation between regression residuals from one period and the previous period.

The test statistic can take on values ranging from 0 to 4.

A value of 2 indicates no serial correlation, a value between 0 and 2 indicates a positive serial correlation, and a value between 2 and 4 indicates a negative serial correlation:

If there is no autocorrelation, the regression errors will be uncorrelated, and thus $DW=2$

$$DW=2(1-r)=2(1-0)=2$$

PREDICTIVE ANALYTICS

For positive serial autocorrelation, $DW < 2$. For example, if serial correlation of the regression residuals = 1, $DW = 2(1-1) = 0$

For negative autocorrelation, $DW > 2$. For example, if serial correlation of the regression residual = -1, $DW = 2(1-(-1)) = 4$.

To reject the null hypothesis of no serial correlation, we need to find a critical value lower than our calculated value of d^* . Unfortunately, we cannot know the true critical value, but we can narrow down the range of possible values.

Define d_l as the lower value and d_u as the upper value:

If the DW statistic is less than d_l , we reject the null hypothesis of no positive serial correlation.

If the DW statistic is greater than $(4-d_l)$, we reject the null hypothesis, indicating a significant negative serial correlation.

If the DW statistic falls between d_l and d_u , the test results are inconclusive.

If the DW statistic is greater than d_u , we fail to reject the null hypothesis of no positive serial correlation.

Example: The Durbin-Watson Test for Serial Correlation

Consider a regression output with two independent variables that generate a DW statistic of 0.654. Assume that the sample size is 15. Test for serial correlation of the error terms at the 5% significance level.

Solution

From the Durbin-Watson table with $n=15$ and $k=2$, we see that $d_l=0.95$ and $d_u=1.54$.

Since $d=0.654 < 0.95=d_l$, we reject the null hypothesis and conclude that there is significant positive autocorrelation.

Consider a regression model with 80 observations and two independent variables. Assume that the correlation between the error term and the first lagged value of the error term is 0.18.

.

Solution

The correct answer is C.

The test statistic is:

$$DW \approx 2(1-r) = 2(1-0.18) = 1.64$$

The critical values from the Durbin Watson table

with $n=80$ and $k=2$ is $d_l=1.59$ and $d_u=1.69$ Because $1.69 > 1.64 > 1.59$,

We determine the test results are inconclusive.

PREDICTIVE ANALYTICS

Lagged Series and Lag Plots: Lagging a time series means to shift its values forward one or more time steps, or equivalently, to shift the times in its index backward one or more steps. In either case, the effect is that the observations in the lagged series will appear to have happened later in time.

	y	y_lag_1	y_lag_2
Date			
1954-07	5.8	NaN	NaN
1954-08	6.0	5.8	NaN
1954-09	6.1	6.0	5.8
1954-10	5.7	6.1	6.0
1954-11	5.3	5.7	6.1

we could use `y_lag_1` and `y_lag_2` as features to predict the target `y`. This would forecast the future unemployment rate as a function of the unemployment rate in the prior two months

Lag feature example

AutoML generates lags with respect to the forecast horizon. The example in this section illustrates this concept. Here, we use a forecast horizon of three and target lag order of one. Consider the following monthly time series:

Table 1: Original time series

Date	??
1/1/2001	0
2/1/2001	10
3/1/2001	20
4/1/2001	30
5/1/2001	40
6/1/2001	50

First, we generate the lag feature for the horizon $h=1$ only. As you continue reading, it will become clear why we use individual horizons in each table.

PREDICTIVE ANALYTICS

Table 2: Lag featurization for $h=1$

Date	$\diamond\diamond$	Origin	$\diamond\diamond-1$	h
1/1/2001	0	12/1/2000	-	1
2/1/2001	10	1/1/2001	0	1
3/1/2001	20	2/1/2001	10	1
4/1/2001	30	3/1/2001	20	1
5/1/2001	40	4/1/2001	30	1
6/1/2001	50	5/1/2001	40	1

Table 2 is generated from Table 1 by shifting the $\diamond\diamond$ column down by a single observation. We've added a column named Origin that has the dates that the lag features originate from. Next, we generate the lagging feature for the forecast horizon $h=2$ only.

Table 3: Lag featurization for $h=2$

Date	$\diamond\diamond$	Origin	$\diamond\diamond-2$	h
1/1/2001	0	11/1/2000	-	2
2/1/2001	10	12/1/2000	-	2
3/1/2001	20	1/1/2001	0	2
4/1/2001	30	2/1/2001	10	2
5/1/2001	40	3/1/2001	20	2
6/1/2001	50	4/1/2001	30	2

Table 3 is generated from Table 1 by shifting the $\diamond\diamond$ column down by two observations. Finally, we will generate the lagging feature for the forecast horizon $h=3$ only.

Table 4: Lag featurization for $h=3$

Date	$\diamond\diamond$	Origin	$\diamond\diamond-3$	h
1/1/2001	0	10/1/2000	-	3
2/1/2001	10	11/1/2000	-	3
3/1/2001	20	12/1/2000	-	3
4/1/2001	30	1/1/2001	0	3
5/1/2001	40	2/1/2001	10	3
6/1/2001	50	3/1/2001	20	3

Next, we concatenate Tables 1, 2, and 3 and rearrange the rows. The result is in the following table:

PREDICTIVE ANALYTICS

Table 5: Lag featurization complete

Date	$yt-1$	Origin	$yt-1(h)$	h
1/1/2001	0	12/1/2000	-	1
1/1/2001	0	11/1/2000	-	2
1/1/2001	0	10/1/2000	-	3
2/1/2001	10	1/1/2001	0	1
2/1/2001	10	12/1/2000	-	2
2/1/2001	10	11/1/2000	-	3
3/1/2001	20	2/1/2001	10	1
3/1/2001	20	1/1/2001	0	2
3/1/2001	20	12/1/2000	-	3
4/1/2001	30	3/1/2001	20	1
4/1/2001	30	2/1/2001	10	2
4/1/2001	30	1/1/2001	0	3
5/1/2001	40	4/1/2001	30	1
5/1/2001	40	3/1/2001	20	2
5/1/2001	40	2/1/2001	10	3
6/1/2001	50	4/1/2001	40	1
6/1/2001	50	4/1/2001	30	2
6/1/2001	50	3/1/2001	20	3

In the final table, we've changed the name of the lag column to $yt-1$ to reflect that the lag is generated with respect to a specific horizon. The table shows that the lags we generated with respect to the horizon can be mapped to the conventional ways of generating lags in the previous tables.

Regression Diagnostics: Diagnostics for regression models are tools that assess a model's compliance to its assumptions and investigate if there is a single observation or group of observations that are not well represented by the model. These tools allow researchers to evaluate if a model appropriately represents the data of their study.

Model Assumptions

The model fitting is just the first part of the story for regression analysis since this is all based on certain assumptions. Regression diagnostics are used to evaluate the model assumptions and investigate whether or not there are observations with a large, undue influence on the analysis. Again, the assumptions for linear regression are:

1. **Linearity:** The relationship between X and the mean of Y is linear.
2. **Homoscedasticity:** The variance of residual is the same for any value of X.
3. **Independence:** Observations are independent of each other.

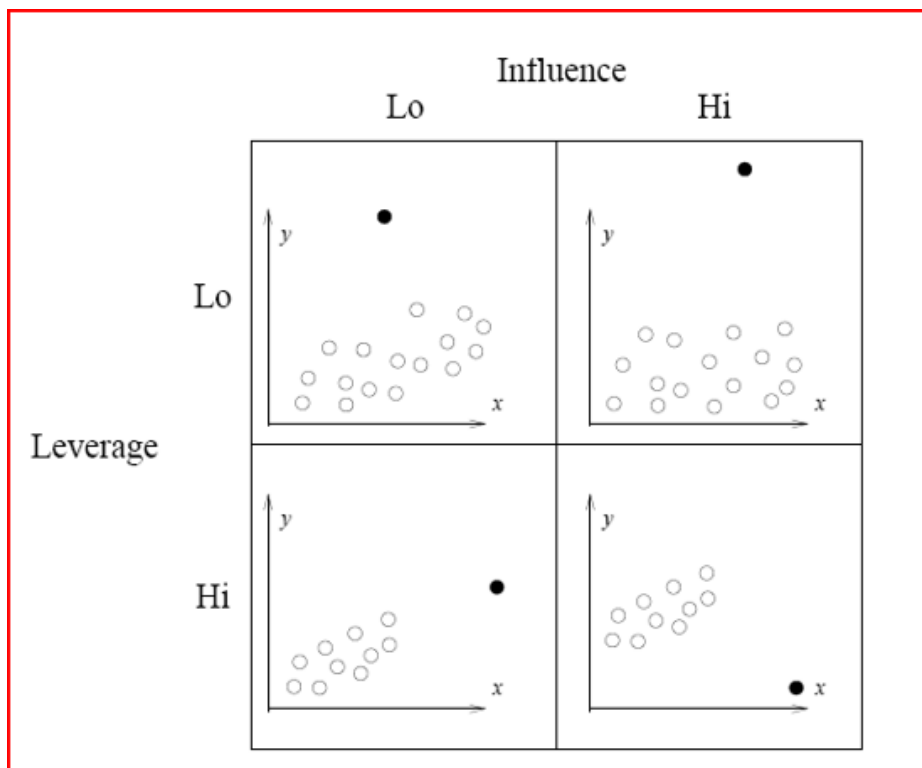
PREDICTIVE ANALYTICS

4. **Normality:** For any fixed value of X, Y is normally distributed.

Before we go further, let's review some definitions for problematic points.

- **Outliers:** an outlier is defined as an observation that has a large residual. In other words, the observed value for the point is very different from that predicted by the regression model.
- **Leverage points:** A leverage point is defined as an observation that has a value of x that is far away from the mean of x.
- **Influential observations:** An influential observation is defined as an observation that changes the slope of the line. Thus, influential points have a large influence on the fit of the model. One method to find influential points is to compare the fit of the model with and without each observation.

Illustration of Influence and leverage



The diagnostic plots show residuals in four different ways. Let's take a look at the first type of plot:

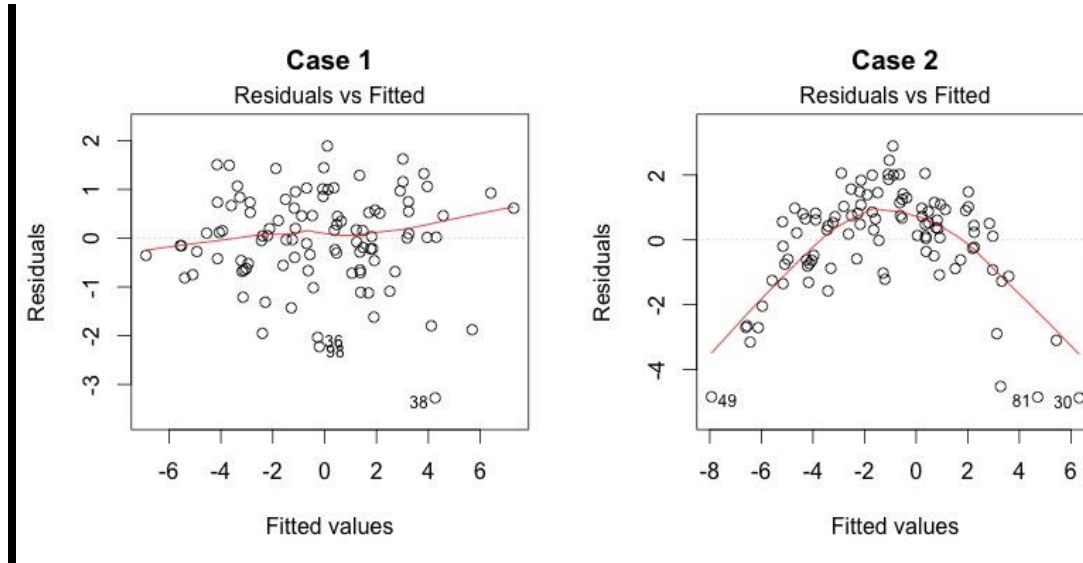
1. Residuals vs Fitted

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable, and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals

PREDICTIVE ANALYTICS

around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

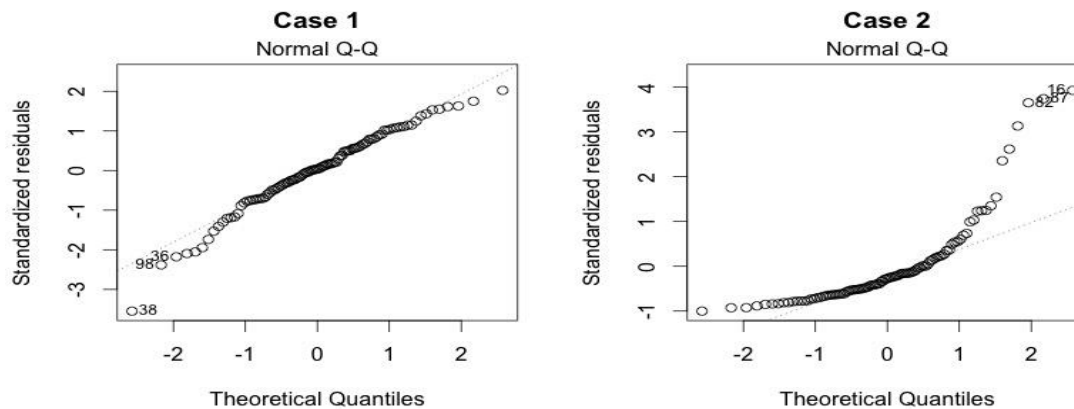
Let's look at residual plots from a 'good' model and a 'bad' model. The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not.



I don't see any distinctive pattern in Case 1, but I see a parabola in Case 2, where the non-linear relationship was not explained by the model and was left out in the residuals.

2. Normal Q-Q

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

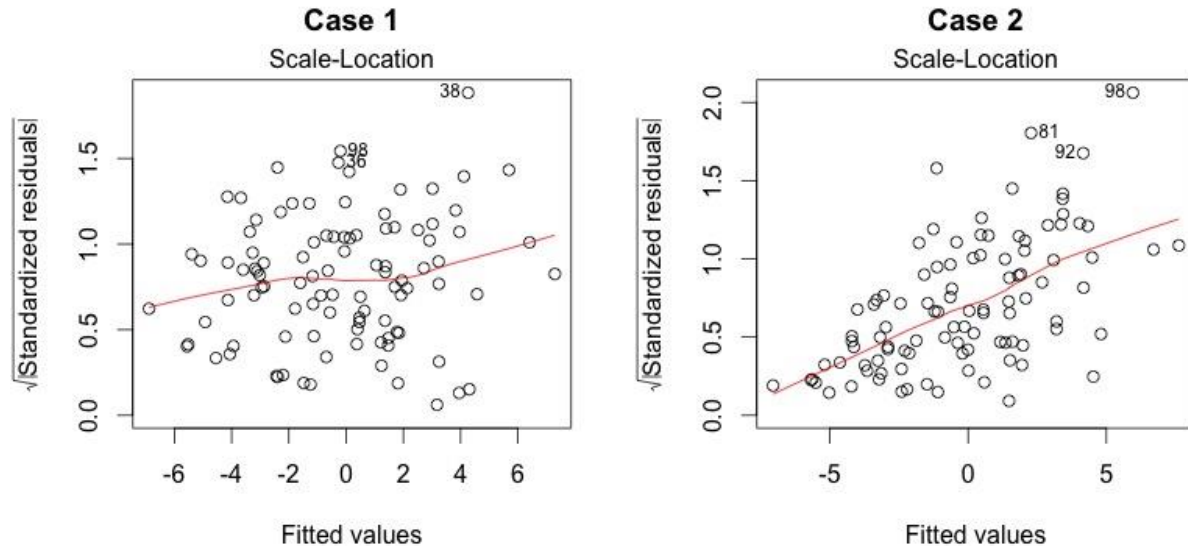


They wouldn't be a perfect straight line, Case 2 definitely concerns

PREDICTIVE ANALYTICS

3. Scale-Location

It's also called a Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.



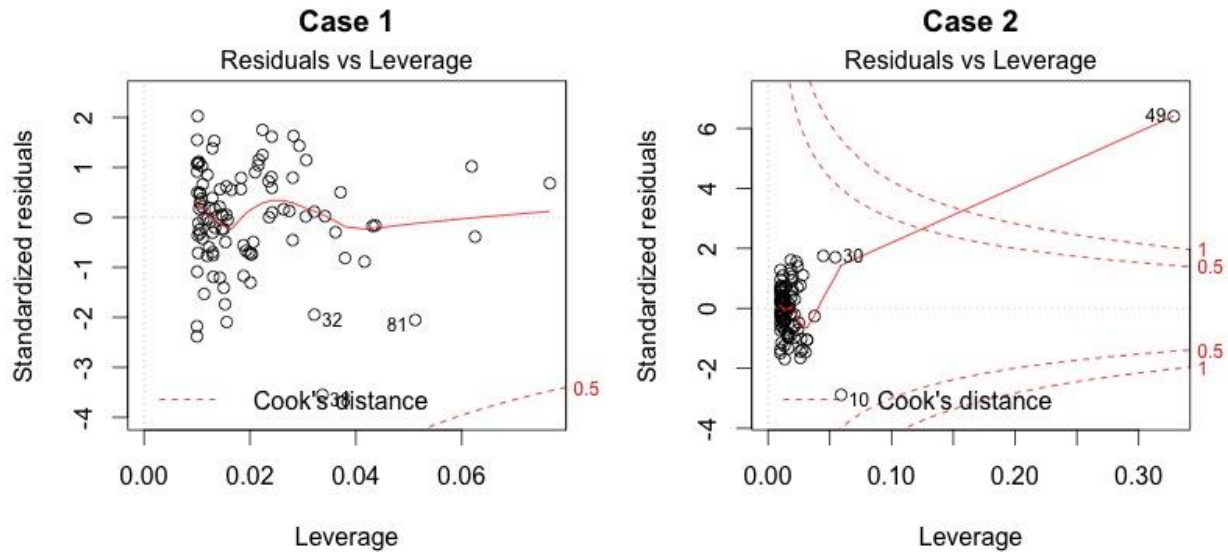
In Case 1, the residuals appear randomly spread, whereas in Case 2, the residuals begin to spread wider along the x-axis as it passes around . Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.

4. Residuals vs Leverage

This plot helps us to find influential cases (i.e., subjects) if there are any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of the dashed lines. When cases are outside of the dashed lines (meaning they have high "Cook's distance" scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

PREDICTIVE ANALYTICS



Case 1 is the typical look when there is no influential case, or cases. You can barely see Cook's distance lines (red dashed lines) because all cases are well inside of the Cook's distance lines. In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot).

The four plots show potential problematic cases with the row numbers of the cases in the data set.

PREDICTIVE ANALYTICS

UNIT-3 DUMMY MODELLING

What is a Dummy variable: A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.

How to Use Dummy Variables in Regression Analysis

Linear regression is a method we can use to quantify the relationship between one or more predictor variables and a **response variable**.

Typically we use linear regression with **quantitative variables**. Sometimes referred to as “numeric” variables, these are variables that represent a measurable quantity. Examples include:

- Number of square feet in a house
- Population size of a city
- Age of an individual

However, sometimes we wish to use categorical variables as predictor variables. These are variables that take on names or labels and can fit into categories. Examples include:

- Eye color (e.g. “blue”, “green”, “brown”)
- Gender (e.g. “male”, “female”)
- Marital status (e.g. “married”, “single”, “divorced”)

When using categorical variables, it doesn’t make sense to just assign values like 1, 2, 3, to values like “blue”, “green”, and “brown” because it doesn’t make sense to say that green is twice as colorful as blue or that brown is three times as colorful as blue.

Instead, the solution is to use **dummy variables**. These are variables that we create specifically for regression analysis that take on one of two values: zero or one.

Dummy Variables: Numeric variables used in regression analysis to represent categorical data that can only take on one of two values: zero or one.

The number of dummy variables we must create is equal to $k-1$ where k is the number of different values that the categorical variable can take on.

The following examples illustrate how to create dummy variables for different datasets.

PREDICTIVE ANALYTICS

Example 1: Create a Dummy Variable with Only Two Values

Suppose we have the following dataset and we would like to use *gender* and *age* to predict *income*:

Income	Age	Gender
\$45,000	23	Male
\$48,000	25	Female
\$54,000	24	Male
\$57,000	29	Female
\$65,000	38	Female
\$69,000	36	Female
\$78,000	40	Male
\$83,000	59	Female
\$98,000	56	Male
\$104,000	64	Male
\$107,000	53	Male

To use *gender* as a predictor variable in a regression model, we must convert it into a dummy variable.

Since it is currently a categorical variable that can take on two different values (“Male” or “Female”), we only need to create $k-1 = 2-1 = 1$ dummy variable.

To create this dummy variable, we can choose one of the values (“Male” or “Female”) to represent 0 and the other to represent 1.

In general, we usually represent the most frequently occurring value with a 0, which would be “Male” in this dataset.

Thus, here’s how we would convert *gender* into a dummy variable:

Income	Age	Gender		Income	Age	Gender_Dummy
\$45,000	23	Male	→	\$45,000	23	0
\$48,000	25	Female		\$48,000	25	1
\$54,000	24	Male		\$54,000	24	0
\$57,000	29	Female		\$57,000	29	1
\$65,000	38	Female		\$65,000	38	1
\$69,000	36	Female		\$69,000	36	1
\$78,000	40	Male		\$78,000	40	0
\$83,000	59	Female		\$83,000	59	1
\$98,000	56	Male		\$98,000	56	0
\$104,000	64	Male		\$104,000	64	0
\$107,000	53	Male		\$107,000	53	0

PREDICTIVE ANALYTICS

We could then use *Age* and *Gender Dummy* as predictor variables in a regression model.

Create a Dummy Variable with Multiple Values

Suppose we have the following dataset and we would like to use *marital status* and *age* to predict *income*:

Income	Age	Marital Status
\$45,000	23	Single
\$48,000	25	Single
\$54,000	24	Single
\$57,000	29	Single
\$65,000	38	Married
\$69,000	36	Single
\$78,000	40	Married
\$83,000	59	Divorced
\$98,000	56	Divorced
\$104,000	64	Married
\$107,000	53	Married

To use *marital status* as a predictor variable in a regression model, we must convert it into a dummy variable.

Since it is currently a categorical variable that can take on three different values (“Single”, “Married”, or “Divorced”), we need to create $k-1 = 3-1 = 2$ dummy variables.

To create this dummy variable, we can let “Single” be our baseline value since it occurs most often. Thus, here’s how we would convert *marital status* into dummy variables:

Income	Age	Marital Status	Married	Divorced
\$45,000	23	Single	0	0
\$48,000	25	Single	0	0
\$54,000	24	Single	0	0
\$57,000	29	Single	0	0
\$65,000	38	Married	1	0
\$69,000	36	Single	0	0
\$78,000	40	Married	1	0
\$83,000	59	Divorced	0	1
\$98,000	56	Divorced	0	1
\$104,000	64	Married	1	0
\$107,000	53	Married	1	0

PREDICTIVE ANALYTICS

We could then use *Age*, *Married*, and *Divorced* as predictor variables in a regression model.

How to Interpret Regression Output with Dummy Variables

Suppose we fit a [multiple linear regression](#) model using the dataset in the previous example with *Age*, *Married*, and *Divorced* as the predictor variables and *Income* as the response variable.

Here's the regression output:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	14276.12	10411.50	1.37	0.21
Age	1471.67	354.44	4.15	0.00
Married	2479.75	9431.26	0.26	0.80
Divorced	-8397.40	12771.36	-0.66	0.53

The fitted regression line is defined as:

$$\text{Income} = 14,276.21 + 1,471.67*(\text{Age}) + 2,479.75*(\text{Married}) - 8,397.40*(\text{Divorced})$$

We can use this equation to find the estimated income for an individual based on their age and marital status. For example, an individual who is 35 years old and married is estimated to have an income of **\$68,264**:

$$\text{Income} = 14,276.21 + 1,471.67*(35) + 2,479.75*(1) - 8,397.40*(0) = \$68,264$$

Here is how to interpret the regression coefficients from the table:

Intercept: The intercept represents the average income for a single individual who is zero years old. Obviously you can't be zero years old, so it doesn't make sense to interpret the intercept by itself in this particular regression model.

Age: Each one year increase in age is associated with an average increase of \$1,471.67 in income. Since the p-value (.00) is less than .05, age is a statistically significant predictor of income.

Married: A married individual, on average, earns \$2,479.75 more than a single individual. Since the p-value (0.80) is not less than .05, this difference is not statistically significant.

Divorced: A divorced individual, on average, earns \$8,397.40 less than a single individual. Since the p-value (0.53) is not less than .05, this difference is not statistically significant.

PREDICTIVE ANALYTICS

Since both dummy variables were not statistically significant, we could drop marital status as a predictor from the model because it doesn't appear to add any predictive value for income.

A **dummy** variable is a binary variable that takes a value of 0 or 1. One adds such variables to a regression model to represent factors which are of a binary nature i.e. they are either observed or not observed.

Several interesting use cases. Here are some of them:

For representing a Yes/No property: To indicate whether a data point has a certain property. For example, a dummy variable can be used to indicate whether a car engine is of type 'Standard' or 'Turbo'. Or if a participant in a drug trial belongs to the placebo group or the treatment group.

For representing a categorical value: A related use of dummies is to indicate which one of a set of categorical values a data point belongs to. For example, a vehicle's body style could be one of convertible, hatchback, coupe, sedan, or wagon. In this case, we would add five dummy variables to the data set, one for each of the 5 body styles and we would 'one hot encode' this five element vector of dummies. Thus, the vector [0, 1, 0, 0, 0] would represent all hatchbacks in the data set.

For representing an ordered categorical value: An extension of the use of dummies to represent categorical data is one where the categories are ordered. Suppose our Automobiles data set contains cars with engines having 2,3,4,5,6,8 or 12 cylinders. Here, we need to also capture the information contained in the ordering.

For representing a seasonal period: A dummy variable can be added to represent each one of the possibly many seasonal periods contained in the data. For example, the flow of traffic through intersections often exhibits seasonality at an hourly level (they are highest during the morning and evening rush hours) and also a weekly period (lowest on Sundays). Adding dummy variables to the data for each of the two seasonal periods will allow you explain away much of the variation in the traffic flow that is attributable to daily and weekly variations.

For representing Fixed Effects: While building regression models for panel data sets, dummies can be used to represent 'unit-specific' and 'time-specific' effects, especially in a Fixed Effects regression model.

For representing Treatment Effects: In a treatment effects model, a dummy variable can be used to represent the effect of both time (i.e. the effect before and after treatment is applied), the effect of group membership (whether the participant received the treatment or the placebo), and the effect of the interaction between the time and group memberships.

PREDICTIVE ANALYTICS

In regression discontinuity designs: This is best explained with an example. Imagine a data set of monthly employment rate numbers that contains a sudden, sharp increase in the unemployment rate caused by a brief and severe recession. For this data, a regression model used for modeling the unemployment rate can deploy a dummy variable to estimate the expected impact of the recession on the unemployment rate.

1. A dummy variable takes on 1 and 0 only. The number 1 and 0 have no numerical (quantitative) meaning. The two numbers are used to represent groups. In short dummy variable is categorical (qualitative).

(a) For instance, we may have a sample (or population) that includes both female and male. Then a dummy variable can be defined as $D = 1$ for female and $D = 0$ for male. Such a dummy variable divides the sample into two subsamples (or two sub-populations): one for female and one for male.

(b) Dummy variable follows Bernoulli distribution. The distribution is characterized by the parameter p

$D = 1$, with probability p

0, with probability $1 - p$

Consider using dummy variable as regressor

$$Y = \beta_0 + \beta_1 D + u$$

Regression can be broken into two separate regressions as

$$Y = \beta_0 + u, \text{ when } D = 0$$

$$(\beta_0 + \beta_1) + u, \text{ when } D = 1$$

Taking expectation leads to $E(Y | D = 0) = \beta_0$

$$E(Y | D = 1) = \beta_0 + \beta_1$$

$$\beta_0 = E(Y | D = 0)$$

$$\beta_1 = E(Y | D = 1) - E(Y | D = 0)$$

Therefore β_0 is the mean of Y conditional on $D = 0$ (or mean of Y in the subpopulation with $D = 0$),

β_1 is the difference in mean Y between the two sub-populations.

PREDICTIVE ANALYTICS

Linear Probability Modeling

Linear Probability Modeling (LPM) is a statistical technique used to model binary dependent variables in regression analysis. It is a simple method that assumes a linear relationship between a set of independent variables and the probability of a binary event occurring, such as whether or not a customer will purchase a product, whether or not a student will pass an exam, or whether or not a patient will respond to a certain treatment.

Applications of LPM

Marketing and Consumer Research: LPM can be used to model the likelihood that a customer will make a purchase, sign up for a subscription, or engage with a brand in some other way. By identifying the factors that influence these behaviors, marketers can optimize their campaigns and improve their return on investment (ROI).

Health Outcomes and Policy: LPM can be used to model the likelihood that a patient will respond to a certain treatment or experience a particular health outcome, such as remission or readmission. This can help inform clinical decision-making, health policy, and resource allocation.

Education and Social Policy: LPM can be used to model the likelihood that a student will pass an exam, graduate from high school, or attend college. This can help identify the factors that contribute to educational success and inform policies aimed at improving educational outcomes and reducing disparities.

Labor Economics and Employment: LPM can be used to model the likelihood that a worker will be employed, experience wage growth, or change occupations. This can help inform policies related to job training, workforce development, and labor market regulation.

Political Science and Public Opinion: LPM can be used to model the likelihood that a voter will support a particular candidate or issue. This can help identify the factors that influence voting behavior and inform political strategies and policy advocacy.

Logit model and applications

A Logit model, also known as logistic regression, is a statistical technique used to analyze the relationship between one or more independent variables and a binary dependent variable. The dependent variable takes on only two possible values, typically coded as 0 or 1, representing a "failure" or a "success", respectively.

Here are some common applications of Logit model:

Business and Marketing: Logit model can be used to model customer behavior, such as whether or not a customer will purchase a product or service. By identifying the factors that influence purchasing decisions, businesses can better target their marketing campaigns and optimize their product offerings.

Medicine and Healthcare: Logit model can be used to model the likelihood of a patient experiencing a certain health outcome, such as whether or not a patient will have a heart attack or develop a certain disease. This can help clinicians to predict and prevent adverse health outcomes and inform clinical decision-making.

Finance and Economics: Logit model can be used to model the likelihood of an event occurring, such as the likelihood of a company defaulting on its debt or the likelihood of a stock market crash. This can help investors and analysts to make informed financial decisions and manage risk.

PREDICTIVE ANALYTICS

Social Sciences: Logit model can be used to model the likelihood of a certain behavior or outcome occurring in a social context, such as the likelihood of an individual engaging in risky behavior or the likelihood of a community experiencing a social issue like homelessness or substance abuse.

Political Science and Public Opinion: Logit model can be used to model voting behavior and public opinion on certain issues. By identifying the factors that influence voting and opinion, policymakers can better understand and respond to the concerns of their constituents.

Probit model and applications

A Probit model is a statistical model that is used to analyze the relationship between one or more independent variables and a binary dependent variable, with the difference being that it assumes that the errors of the dependent variable follow a normal distribution rather than a logistic distribution.

Here are some common applications of Probit model:

Economics and Finance: Probit model can be used to model financial decision-making, such as the likelihood of a borrower defaulting on a loan or the likelihood of an investor making a certain investment. This can help investors and financial institutions to better understand and manage risk.

Marketing and Consumer Research: Probit model can be used to model consumer behavior, such as the likelihood of a customer making a purchase or choosing a certain brand. This can help businesses to better target their marketing campaigns and optimize their product offerings.

Health Outcomes and Policy: Probit model can be used to model the likelihood of a patient experiencing a certain health outcome, such as the likelihood of a patient developing a certain disease. This can help clinicians to predict and prevent adverse health outcomes and inform clinical decision-making.

Environmental and Agricultural Sciences: Probit model can be used to model the likelihood of a certain outcome occurring in a natural environment, such as the likelihood of a species becoming extinct or the likelihood of a crop being affected by a certain pest. This can help policymakers to design and implement effective environmental and agricultural policies.

Social Sciences: Probit model can be used to model the likelihood of a certain behavior or outcome occurring in a social context, such as the likelihood of an individual engaging in risky behavior or the likelihood of a community experiencing a social issue like poverty or inequality.

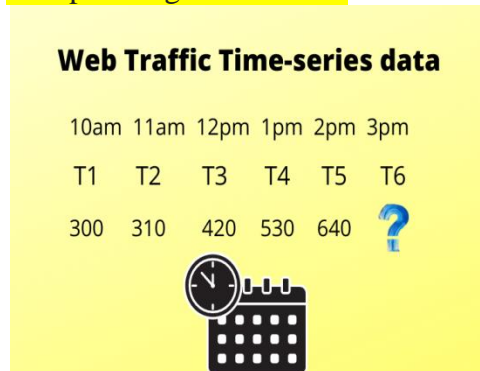
PREDICTIVE ANALYTICS

FORECASTING AND MACHINE LEARNING

A time series is a sequence of observations recorded over a certain period of time. A simple example of time series is how we come across different temperature changes day by day or in a month.

Basics of Time-Series Forecasting

Timeseries forecasting in simple words means to forecast or to predict the future value(eg-stock price) over a period of time. There are different approaches to predict the value, consider an example there is a company XYZ records the website traffic in each hour and now wants to forecast the total traffic of the coming hour. If I ask you what will your approach to forecasting the upcoming hour traffic?



A different person can have a different perspective like one can say find the mean of all observations, one can have like take mean of recent two observations, one can say like give more weightage to current observation and less to past, or one can say use interpolation. There are different methods to forecast the values.

while Forecasting time series values, 3 important terms need to be taken care of and the main task of time series forecasting is to forecast these three terms.

1) Seasonality

Seasonality is a simple term that means while predicting a time series data there are some months in a particular domain where the output value is at a peak as compared to other months. for example if you observe the data of tours and travels companies of past 3 years then you can see that in November and December the distribution will be very high due to holiday season and festival season. So while forecasting time series data we need to capture this seasonality.

2) Trend

The trend is also one of the important factors which describe that there is certainly increasing or decreasing trend time series, which actually means the value of organization or sales over a period of time and seasonality is increasing or decreasing.

3) Unexpected Events

Unexpected events mean some dynamic changes occur in an organization, or in the market which cannot be captured. for example a current pandemic we are suffering from, and if you observe the Sensex or nifty chart there is a huge decrease in stock price which is an unexpected event that occurs in the surrounding.

PREDICTIVE ANALYTICS

Additive and Multiplicative Time series

In the real world, we meet with different kinds of time series data. For this, we must know the concepts of Exponential smoothing and for this first, we need to study types of time series data as additive and multiplicative. As we studied there are 3 components we need to capture as Trend(T), seasonality(S), and Irregularity(I).

Additive time series is a combination(addition) of trend, seasonality, and Irregularity while multiplicative time series is the multiplication of these three terms.

Additive Time Series

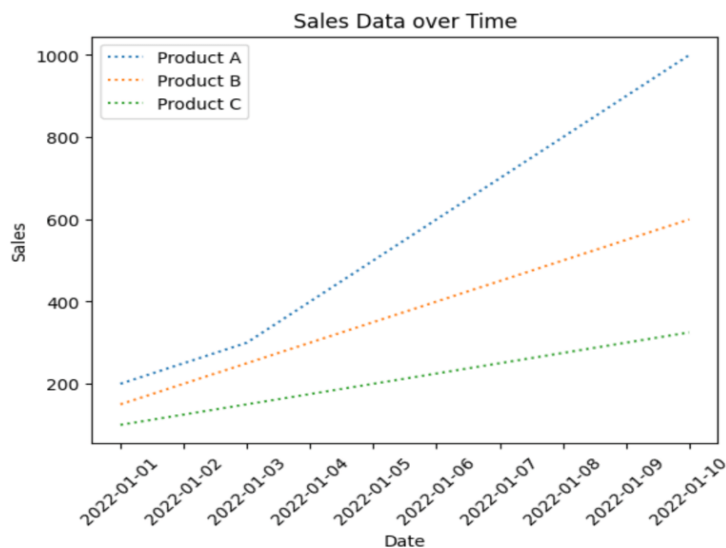
$$\text{Value} = \text{Base Level} + \text{Trend} + \text{Seasonality} + \text{Error}$$

Multiplicative Time Series

$$\text{Value} = \text{Base Level} * \text{Trend} * \text{Seasonality} * \text{Error}$$

Types of Time series plots:

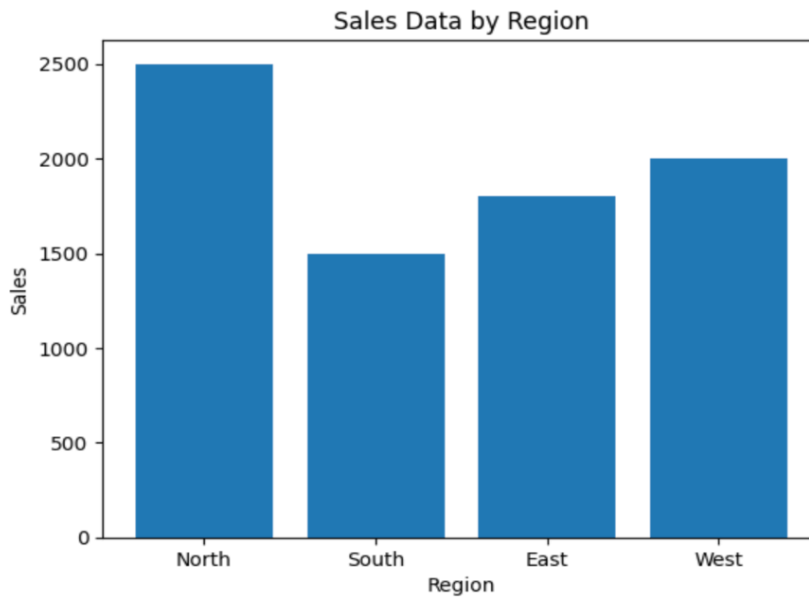
Quantitative variables have numerical values and can be quantified/measured on a discrete or continuous scale. For instance, temperature readings, stock prices, and website traffic data are quantitative variables. They are typically plotted on the y-axis of a time-series chart, with time on the x-axis



Qualitative variables, on the other hand, have categorical or non-numerical values. For instance, a website's name, the type of product sold, or the region where a company operates are qualitative variables. Typically, they are not plotted on a time-series chart as they do not have a numerical

PREDICTIVE ANALYTICS

value that can be represented on a continuous scale. However, we can use them to categorize/group the data for analysis.



To interpret a time-series plot, you must understand the data patterns over time. These are the key factors to consider when interpreting a time-series plot:

Seasonality: It refers to recurring patterns/cycles that occur over a particular period. The patterns can be weekly, monthly, quarterly, or annual. For instance, winter coat sales typically show a seasonal pattern as they increase in the fall and winter and decrease in the spring and summer.

Trend: It refers to the general direction the data moves in over time. A trend can either be upward, downward or remain constant. A positive trend indicates that the values are increasing over time, while a negative trend shows the values are decreasing over time. A horizontal trend suggests that the values remain constant over time.

Outliers: They refer to values that lie outside the usual pattern of the data. External factors or random events can cause outliers and can significantly impact the interpretation of the time-series plot.

Level: It refers to the average data value over the entire time period in a time-series plot. A higher level indicates that the values are higher overall and vice versa.

Trend Analysis?

Trend analysis is a technique used in [technical analysis](#) that attempts to predict future stock price movements based on recently observed trend data.

PREDICTIVE ANALYTICS

A [trend](#) is a general direction the market is taking during a specified period of time. Trends can be both upward and downward, relating to bullish and [bearish markets](#), respectively. While there is no specified minimum amount of time required for a direction to be considered a trend, the longer the direction is maintained, the more notable the trend.

There are three main types of market trend for analysts to consider:

1. **Upward trend:** An upward trend, also known as a [bull market](#), is a sustained period of rising prices in a particular security or market. Upward trends are generally seen as a sign of economic strength and can be driven by factors such as strong demand, rising profits, and favorable economic conditions.
2. **Downward trend:** A downward trend, also known as a [bear market](#), is a sustained period of falling prices in a particular security or market. Downward trends are generally seen as a sign of economic weakness and can be driven by factors such as weak demand, declining profits, and unfavorable economic conditions.
3. **Sideways trend:** A [sideways trend](#), also known as a [rangebound](#) market, is a period of relatively stable prices in a particular security or market. Sideways trends can be characterized by a lack of clear direction, with prices fluctuating within a relatively narrow range.

Trend Trading Strategies

[Trend traders](#) attempt to isolate and extract profit from trends. There are many different trend trading strategies using a variety of technical [indicators](#):

- **Moving Averages:** These strategies involve entering into long positions when a short-term [moving average](#) crosses above a long-term moving [average](#), and entering short positions when a short-term moving average crosses below a long-term moving average.
- **Momentum Indicators:** These strategies involve entering into [long positions](#) when a security is trending with strong momentum and exiting long positions when a security loses momentum. Often, the [relative strength index](#) (RSI) is used in these strategies.
- **Trendlines & Chart Patterns:** These strategies involve entering long positions when a security is trending higher and placing a [stop-loss](#) below key trendline [support levels](#). If the stock starts to reverse, the position is exited for a profit.

Advantages and Disadvantages of Trend Analysis

Advantages

Trend analysis can offer several advantages for investors and traders. It is a powerful tool for investors and traders as it can help identify opportunities for buying or selling securities, minimize risk, improve decision-making, and enhance portfolio performance.

Trend analysis can be based on a variety of data points, including financial statements, economic indicators, and market data, and there are several different methods that can be used to analyze trends, including technical analysis and fundamental analysis. By providing a deeper understanding of the factors that are driving trends in data, trend analysis can help investors and traders make more informed and confident decisions about their investments.

Disadvantages

Trend analysis can have some potential disadvantages as a tool for making investment decisions. One of these disadvantages is that the accuracy of the analysis depends on the quality of the data being used. If the data is incomplete, inaccurate, or otherwise flawed, the analysis may be misleading or inaccurate.

Another potential disadvantage is that trend analysis is based on historical data, which means it can only provide a limited perspective on the future. While trends in data can provide useful insights, it's important to remember that the future is not necessarily predetermined by the past, and unexpected events or changes in market conditions can disrupt trends. Trend analysis is also focused on identifying patterns in data over a given period of time, which means it may not consider other important factors that could impact the performance of a security or market.

Applications of Time Series Analysis:

1. Time series in Financial and Business Domain

Most financial, investment and business decisions are taken into consideration on the basis of future changes and demands forecasts in the financial domain.

Time series analysis and forecasting essential processes for explaining the dynamic and influential behaviour of financial markets. Via examining financial data, an expert can predict required forecasts for important financial applications in several areas such as risk evolution, [option pricing & trading](#), portfolio construction, etc.

For example, time series analysis has become the intrinsic part of [financial analysis](#) and can be used in predicting interest rates, foreign currency risk, volatility in stock markets and many more. Policymakers and business experts use financial forecasting to make decisions about production, purchases, market sustainability, allocation of resources, etc.

In investment, this analysis is employed to track the price fluctuations and price of a security over time. For instance, the price of a security can be recorded;

- For the short term, such as the observation per hour for a business day, and
- For the long term, such as observation at the month end for five years.

Time series analysis is extremely useful to observe how a given asset, security, or economic variable behaves/changes over time. For example, it can be deployed to evaluate how the underlying changes associated with some data observation behave after shifting to other data observations in the same time period.

2. Time series in Medical Domain

PREDICTIVE ANALYTICS

Medicine has evolved as a data-driven field and continues to contribute in time series analysis to human knowledge with enormous developments.

Case study

Consider the case of combining time series with a medical method CBR (case-based reasoning) and data mining, these synergies are essential as the pre-processing for feature mining from time series data and can be useful to study the progress of patients over time.

In the medical domain, it is important to examine the transformation of behaviour over time as compared to derive inferences depending on the absolute values in the time series. For example, to diagnose heart rate variability in occurrence with respiration based on the sensor readings is the characteristic illustration of connecting time series with case-based monitoring.

However, time series in the context of the epidemiology domain has emerged very recently and incrementally as time series analysis approaches demand recordkeeping systems such that records should be connected over time and collected precisely at regular intervals.

As soon as the government has placed sufficient scientific instruments to accumulate good and lengthy temporal data, healthcare applications using time series analysis have resulted in huge prognostication for the industry as well as for individuals' health diagnoses.

Medical Instruments

Time series analysis has made its way into medicine with the advent of medical devices such as

- Electrocardiograms (ECGs), invented in 1901: For diagnosing cardiac conditions by recording the electrical pulses passing through the heart.
- Electroencephalogram (EEG), invented in 1924: For measuring electrical activity/impulses in the brain.

These inventions made more opportunities for medical practitioners to deploy time series for medical diagnosis.

With the advent of wearable sensors and smart electronic healthcare devices, now persons can take regular measurements automatically with minimal inputs, resulting in a good collection of longitudinal medical data for both sick and healthy individuals consistently.

3. Time Series in Astronomy

One of the contemporary and modern applications where time series plays a significant role are different areas of astronomy and astrophysics,

Being specific in its domain, astronomy hugely relies on plotting objects, trajectories and accurate measurements, and due to the same, astronomical experts are proficient in time series in calibrating instruments and studying objects of their interest.

Time series data had an intrinsic impact on knowing and measuring anything about the universe, it has a long history in the astronomy domain, for example, sunspot time series were recorded in China in 800 BC, which made sunspot data collection as well-recorded natural phenomena.

Similarly, in past centuries, time series analysis was used

- To discover variable stars that are used to surmise stellar distances, and
- To observe transitory events such as supernovae to understand the mechanism of the changing of the universe with time.

Such mechanisms are the results of constant monitoring of live streaming of time series data depending upon the wavelengths and intensities of light that allows astronomers to catch events as they are occurring.

In the last few decades, data-driven astronomy introduced novel areas of research as astroinformatics and astrostatistics; these paradigms involve major disciplines such as [statistics](#), data mining, machine learning and computational intelligence. And here, the role of time series analysis would be detecting and classifying astronomical objects swiftly along with the characterization of novel phenomena independently.

4. Time series in Forecasting Weather

Anciently, the Greek philosopher Aristotle researched weather phenomena with the idea to identify causes and effects in weather changes. Later on, scientists started to accumulate weather-related data using the instrument “barometer” to compute the state of atmospheric conditions, they recorded weather-related data on intervals of hourly or daily basis and kept them in different locations.

With the time, customized weather forecasts began printed in newspapers and later on with the advancement in technology, currently forecasts are beyond the general weather conditions.

In order to conduct atmospheric measurements with computational methods for fast compilations, many governments have established thousands of weather forecasting stations around the world.

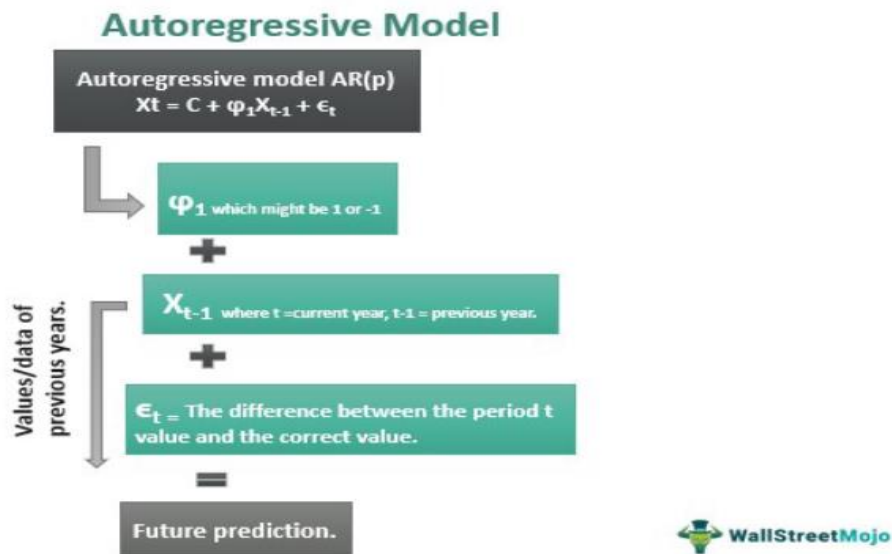
These stations are equipped with highly functional devices and are interconnected with each other to accumulate weather data at different geographical locations and forecast weather conditions at every bit of time as per requirements.

5. Time series in Business Development

Time series forecasting helps businesses to make informed business decisions, as the process analyzes past data patterns it can be useful in forecasting future possibilities and events in the following ways;

- **Reliability:** When the data incorporates a broad spectrum of time intervals in the form of massive observations for a longer time period, time series forecasting is highly reliable. It provides elucidate information by exploiting data observations at various time intervals.
- **Growth:** In order to evaluate the overall financial performance and growth as well as endogenous, time series is the most suitable asset. Basically, endogenous growth is the progress within organizations' internal human capital resulting in [economic growth](#). For example, studying the impact of any policy variables can be manifested by applying time series forecasting.
- **Trend estimation:** Time series methods can be conducted to discover trends, for example, these methods inspect data observations to identify when measurements reflect a decrease or increase in sales of a particular product.
- **Seasonal patterns:** Recorded data points variances could unveil seasonal patterns & fluctuations that act as a base for data forecasting. The obtained information is significant for markets whose products fluctuate seasonally and assist organizations in planning product development and delivery requirements.

An autoregressive model is a process used to predict the future based on accumulated data from the past. It is possible because there is a correlation between the two. Such a model can represent any random procedure where the output is dependent on any previous values.



PREDICTIVE ANALYTICS

This model is often used to predict the future trend in [stock](#) prices by analyzing past performance. Thus, it assumes that the future result will be similar to the previous years. However, this is only sometimes acceptable because, due to continuous global technological and economic changes, there is no guarantee that the future will reflect the past.

The autoregressive (AR) model predicts the future based on past data or information. It helps in stock price forecasting on the assumption that the prices of previous years will genuinely reflect the end.

In an autoregressive model [time series](#) is calculated based on the correlation of past and future data. So, it is a statistical method for any fundamental or [technical analysis](#). But the downside of this model is the assumption that all forces or factors that affected past performance will remain the same, which is unrealistic since change is inevitable in all fields. There is a rapid transformation all around due to endless innovation taking place.

A vector autoregressive model, for instance, consists of multiple variables that attempt to correlate a variable's present values with its past values and the system's past data of other variables. Thus, it is a multivariate model. If an AR model is univariate, it is impossible to get a two-way result between the variables.

The use of the autoregressive process to make forecasts is very significant. However, these models are also stochastic, meaning they have an element of uncertainty. Any unforeseen contingency or sudden change and shift in the [economy](#) will affect the outcome of future values significantly, which refers to the fact that the result will never be accurate. Nevertheless, it is possible to get the closest possible outcome.

A first-order autoregressive model assumes that the immediately previous value decides the current value. However, there might be cases that the present value will depend on two previous values. Thus, in an autoregressive model, time series plays an important role and is used depending on the situation and desired result.

Formula

Let us try to understand the autoregressive model equation as mentioned below:

In this model, some specific values of X_t serve as variables. They have lagged values, which means the past or current output will affect future outcomes.

The autoregressive model equation, denoted by AR(p), is given below:

$$X_t = C + \phi_1 X_{t-1} + \epsilon_t$$

where,

- X_{t-1} = value of X in the previous year/month/week. If “t” is the current year, then “t-1” will be the last.
- ϕ_1 = coefficient, which we multiply with X_{t-1} . The value of ϕ_1 will always be 1 or -1.
- ϵ_t = The difference between the period t value and the correct value ($\epsilon_t = y_t - \hat{y}_t$)
- p = The order. Thus, AR (1) is first order autoregressive model. The second and third order would be AR (2) and AR (3), respectively.

Examples

Here we look at some examples to understand the concept.

Example #1

John is an investor in the [stock market](#). He analyses stocks based on past data related to the company’s performance and [statistics](#). John believes that the performance of the stocks in the previous years strongly correlates with the future, which is beneficial to making [investment](#) decisions.

He uses an autoregressive model with price data for the previous five years. The result gives him an estimate for future prices depending on the assumption that sellers and buyers follow the market movements and accordingly make [investment decisions](#).

Example #2

The concept of AR models has gained [importance](#) in the information technology field. Google has proposed Autoregressive Diffusion Models (ARDMs), which encompass and generalizes the models that depend on any data arrangement. It is possible to train the model to achieve any desired result. Thus, this method will generate outcomes under any order.

Example #3

The autoregression [process](#) can be helpful in the veterinary field; also, the main focus is on the occurrence of a disease over time. In this case, the primary source of information is the systems used to monitor and track the details of animal disease. This data is analyzed and correlated using the model to understand the possibility of any disease occurrence. However, this model has limited use in the veterinary field due to limited data availability and the need for useful software to generate the best results.

Autoregressive Model vs Moving Average

PREDICTIVE ANALYTICS

An autoregressive model is a method of future prediction based on past information. In contrast, a [moving average](#) analyzes data by calculating a series of averages from a large data set. However, the difference between them is as follows:

Autoregressive Model	Moving Average
Use of past data as input.	Use of past errors as input.
The various time slots impact the time.	Some external factors affect the period.
It calculates the regression of past time series.	It calculates the residuals or errors of past time series.
It puts data from the previous time in the regression equation to get the next value.	It states that the next value will be the average of all the past values.
The correlation between the objects of the time series decreases as the time gap increases.	The correlation between the objects of the time series at different points in time is zero.

Predictive analytics using machine learning

Predictive analytics is a powerful technique that ‘predicts’ the future, in a sense. It can help answer key questions, such as how many products a business could sell in the next three months and how much profit it is likely to make.

Using sales as an example, it’s essential to know past sales data in order to predict future sales. The past sales data and cleaned data from descriptive analytics are mixed to create a dataset to train an ML model.

The built model predicts future sales, say, for the next few months. The predicted quantities sold and profits made are compared with the actual numbers sold and profits made. The actual profits could be more or less than what was predicted. The model is refined to overcome such limitations and improve the accuracy of predictions.

Types of analytics

There are four types of analytics: descriptive, diagnostic, predictive, and prescriptive.

PREDICTIVE ANALYTICS

- *Descriptive analytics* deals with the cleaning, relating, summarizing, and visualizing of given data to identify patterns.
- *Diagnostic analytics* deals with analyzing why something is happening. For example, investigating the reason behind the decline or growth of revenue.
- *Predictive analytics* involves predicting future outcomes or unknown events using machine learning and statistical algorithms.
- *Prescriptive analytics* uses descriptive and predictive sources to assist with decision-making.

There are many scenarios where there may be an abundance of data. However, there may be no algorithms available to train machines to perform certain tasks. In this case, we want the machines to learn from the data and apply the learning to unseen inputs. This is referred to as machine learning. For example, if we want to know employee churn rate, we can use a [machine learning model](#) that has been trained on past data to predict if an employee will leave.

ML is used when we cannot explicitly estimate all the possible cases of an event occurring and write a piece of code for each. For example, what are the rules to predict if content posted on a video-hosting platform is for kids or adults? How do we predict the genre of a new show? There are millions of videos uploaded every day. Examining and analyzing each of them manually is impossible. This is where ML comes into play as the algorithms can process enormous amounts of structured (data in rows and columns) and unstructured (images, videos, text with emoticons, etc.) data.

PREDICTIVE ANALYTICS

Steps for predictive analytics using machine learning



There are eight steps to perform predictive analytics with ML.

Step 1: Define the problem statement

We begin by understanding and defining the problem statement, and deciding on the required datasets on which to perform predictive analytics.

Example: There is a grocery store. Our objective is to predict the sales of groceries for the next six months. Here, past sales data of how many groceries were sold and the resulting profits of the last five years will be the dataset.

Step 2: Collect the data

Once we know what sort of dataset is needed to perform predictive analytics using machine learning, we gather all the necessary details that constitute the dataset. We need to ensure that the historical data is collected from an authorized source.

PREDICTIVE ANALYTICS

Using the grocery store example, we can ask the accountant for records of past sales logged in worksheets or billing software. We collect data spanning the past five years.

Step 3: Clean the data

The raw dataset obtained will have some missing data, redundancies, and errors. Since we cannot train the model for predictive analytics directly with such noisy data, we need to clean it. Known as preprocessing, this step involves refining the dataset by eradicating unnecessary and duplicate data.

Step 4: Perform Exploratory Data Analysis (EDA)

EDA involves exploring the dataset thoroughly in order to identify trends, discover anomalies, and check assumptions. It summarizes a dataset's main characteristics. It often uses data visualization techniques.

Step 5: Build a predictive model

Based on the patterns observed in step 4, we build a predictive statistical machine learning model, trained with the cleaned dataset obtained after step 3. This machine learning algorithm helps us perform predictive analytics to foresee the future of our grocery store business. The model can be implemented using [Python](#), R, or MATLAB.

- **Hypothesis testing**

Hypothesis testing can be performed using a standard statistical model. It includes two hypotheses, null and alternate. We either reject or fail to reject the null hypothesis.

Example: A new 'buy one, get one free' scheme is implemented where customers buy a packet of soap and get a face wash for free. Consider the two cases below:

Case 1: Despite the scheme, sales of soap did not improve.

Case 2: After the scheme, sales of soap improved.

If the first case is true, we fail to reject the null hypothesis as there is no improvement. If the second case is true, we reject the null hypothesis.

Step 6: Validate the model

This is a crucial step wherein we check the efficiency of the model by testing it with unseen input datasets. Depending on the extent to which it makes correct predictions, the model is retrained and evaluated.

Step 7: Deploy the model

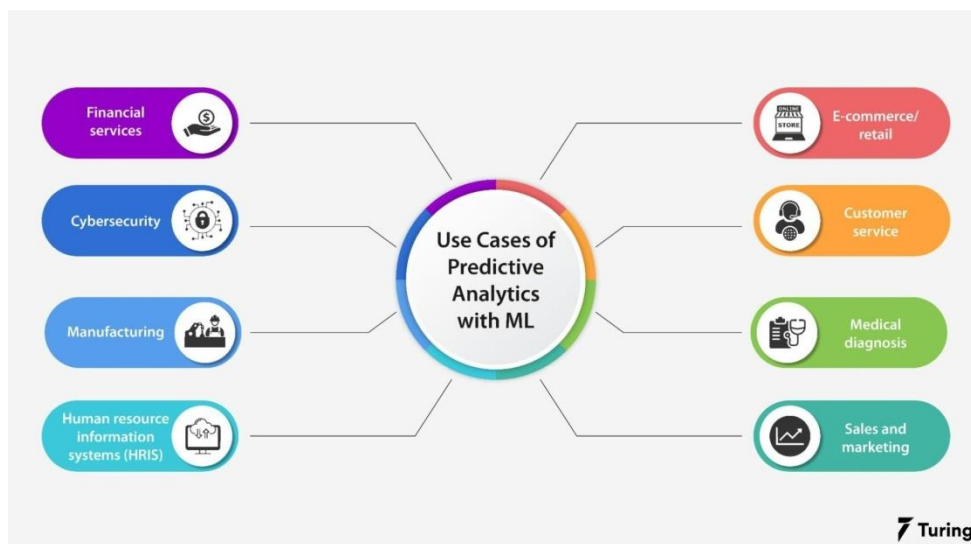
PREDICTIVE ANALYTICS

The model is made available for use in a real-world environment by deploying it on a cloud computing platform so that users can utilize it. Here, the model will make predictions on real-time inputs from the users.

Step 8: Monitor the model

Now that the model is functioning in the real world, we need to verify its performance. Model monitoring refers to examining how the model predicts actual datasets. If any improvement must be made, the dataset is expanded and the model is rebuilt and redeployed.

How machine learning improves predictive analytics



Predictive analytics continues to be improved with machine learning algorithms. The eight use cases discussed below illustrate how.

E-commerce/retail

Predictive analytics achieved through machine learning helps retailers understand customers' preferences. It works by analyzing users' browsing patterns and how frequently a product is clicked on in a website. For example, when we purchase a t-shirt on an e-commerce site, similar shirts are suggested the next time we log in. Sometimes, we may be recommended several specific items that are often purchased together for x amount of money. Such personalized recommendations help retailers retain customers. Predictive analytics also helps maintain inventory by foreseeing and informing sellers about stockouts.

Customer service

Customer segmentation is performed based on insights by predictive analytics. Customers are placed into different segments depending on their purchase patterns. For example, book buyers will form one cluster while t-shirt buyers will constitute another. Tailored marketing strategies are then developed for each of the segments depending on their characteristics.

PREDICTIVE ANALYTICS

Predictive analytics using machine learning can also detect dissatisfied customers and help sellers design products aimed to retain existing customers and attract new ones.

Medical diagnosis

Machine learning models that are trained on large and varied datasets can study patient symptoms comprehensively to provide faster and more accurate diagnoses. Performing predictive analytics on the reasons behind past hospital readmissions can also improve care.

Further, hospitals can use predictive analytics to provide the best care by pre-determining increase of hospital bed availability or staff shortage. For example, if the number of COVID cases for the next month can be predicted and the rise in the number of severely infected can be forecasted, hospitals can make arrangements to deal with such a scenario more efficiently.

Sales and marketing

Predictive analytics of historical data of customer behavior and market trends can help businesses understand the demands of prospective customers. Companies can achieve higher targets by streamlining their sales and marketing activities into a data-based undertaking. Demand forecasting also helps businesses estimate the demand for certain products in the future.

Financial services

Predictive analytics using machine learning helps detect fraudulent activities in the financial sector. Fraudulent transactions are identified by training machine learning algorithms with past datasets. The models find risky patterns in these datasets and learn to predict and deter fraud.

Cybersecurity

[Machine learning](#) algorithms can analyze web traffic in real-time. When an unusual pattern is observed, advanced statistical methods of predictive analytics foresee and prevent cyber-attacks. They also automatically collect attack-related data and generate useful reports on a cyber-attack, thereby reducing the need for manpower.

Manufacturing

Machine learning and predictive analytics help manufacturers monitor machines and notify them when crucial components need to be repaired or replaced. They can also predict market fluctuations, reduce the number of accidents, improve key performance indicators (KPIs), and enhance overall production quality.

Human Resource Information Systems (HRIS)

Predictive analytics using machine learning identifies employee churn rate and keeps human resources (HR) departments informed of the same. Models can be trained with datasets that have details such as an employee's monthly income, allowances, increments, insurance, and so on.

PREDICTIVE ANALYTICS

The models learn from past records of ex-employees and find patterns to understand the reasons for leaving. They then predict if new employees are likely to resign or not, empowering HR to minimize the risk.

Predictive analytics tends to be performed with three main types of techniques:

Regression analysis

Regression is a statistical analysis technique that estimates relationships between variables. Regression is useful to determine patterns in large datasets to determine the correlation between inputs. It is best employed on continuous data that follows a known distribution. Regression is often used to determine how one or more independent variables affects another, such as how a price increase will affect the sale of a product.

Decision trees

Decision trees are classification models that place data into different categories based on distinct variables. The method is best used when trying to understand an individual's decisions. The model looks like a tree, with each branch representing a potential choice, with the leaf of the branch representing the result of the decision. Decision trees are typically easy to understand and work well when a dataset has several missing variables.

Neural networks

Neural networks are machine learning methods that are useful in predictive analytics when modeling very complex relationships. Essentially, they are powerhouse pattern recognition engines. Neural networks are best used to determine nonlinear relationships in datasets, especially when no known mathematical formula exists to analyze the data. Neural networks can be used to validate the results of decision trees and regression models.

Every business seeks to grow. But only a handful of companies that successfully actualize this vision do so through data-based decision making. And to make these informed decisions, companies have been using machine learning-based predictive analytics.

Predictive analytics is predicting future outcomes based on historical and current data. It uses various statistical and data modeling techniques to analyze past data, identify trends, and help make informed business decisions. While previously, machine learning and predictive analytics were viewed as two entirely different and unrelated concepts, the increasing demands of effective data analytics have brought machine learning algorithms to intertwine with predictive analytics. Today, predictive analytics extensively uses machine learning for data modeling due to its ability to accurately process vast amounts of data and recognize patterns.

In this piece, we'll learn in detail how machine learning analytics is helping companies predict the future and make informed decisions.

Predictive Analytics & Machine Learning

1 Predictive analysis is a forward-gazing technique of analyzing historical data to forecast accurate future outcomes based on a variety of set parameters.

2 The increasing demands of effective data analytics have brought machine learning algorithms to intertwine with predictive analytics.

3 Using machine learning algorithms, businesses can optimize and uncover new statistical patterns which form the backbone of predictive analytics.

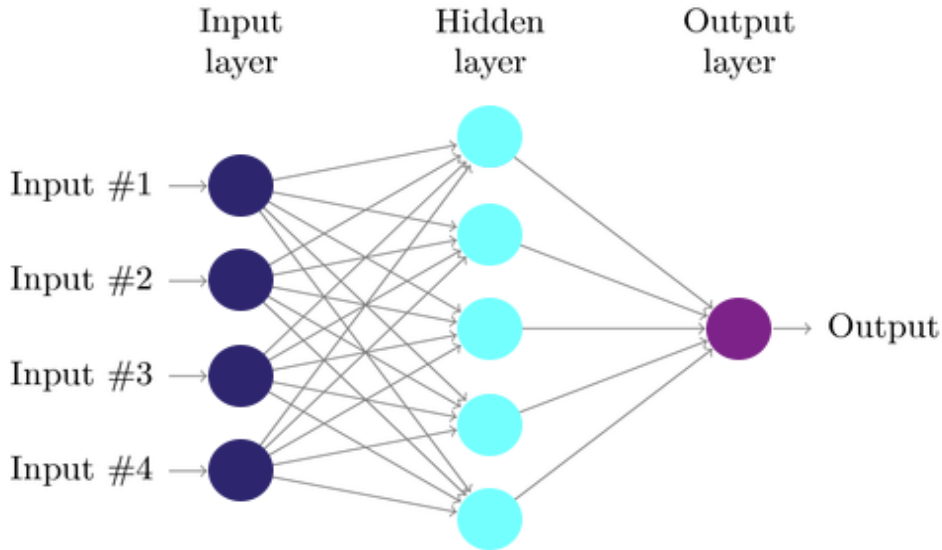
4 Companies are employing machine learning based predictive analytics to gain an edge over the rest of the market.

Copyright © 2020 Maruti Techlabs Inc.

Neural Networks – Building Blocks Of Data Analysis

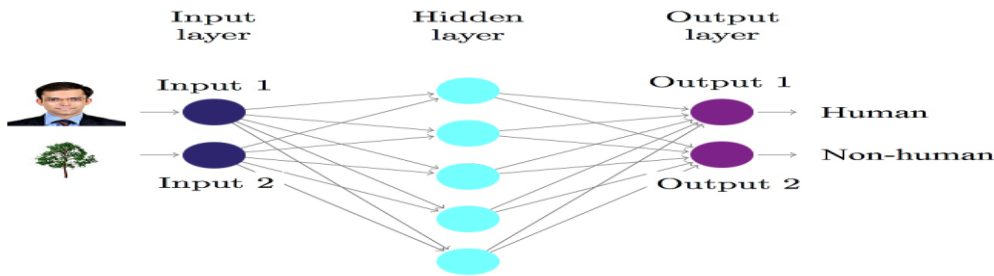
The neural network is a system of hardware and software mimicked after the central nervous system of humans, to estimate functions that depend on vast amounts of unknown inputs. [Neural networks](#) are specified by three things – architecture, activity rule, and learning rule.

According to Kaz Sato, Staff Developer Advocate at Google Cloud Platform, “A neural network is a function that learns the expected output for a given input from training datasets”. A neural network is an interconnected group of nodes. Each processing node has its small sphere of knowledge, including what it has seen and any rules it was initially programmed with or developed for itself.



Neural networks – Building blocks of Data Analysis

In short neural networks are adaptive and modify themselves as they learn from subsequent inputs. For example, below is a representation of a neural network that performs [image recognition](#) for ‘humans’. The system has been trained with a lot of samples of human and non-human images. The resulting network works as a function that takes an image as input and outputs label human or non-human.



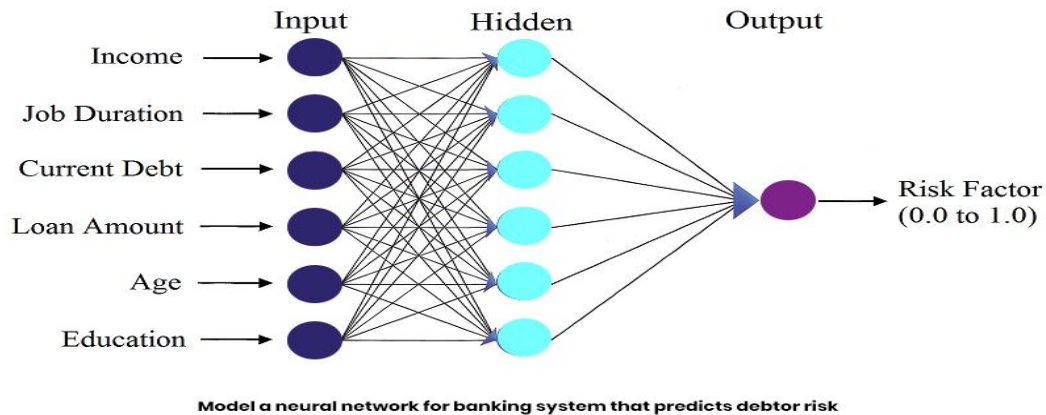
Model for Neural network Image recognition for human and non-human

Building Predictive Capabilities Using Machine Learning And Artificial Intelligence

Let’s implement what we have learned about neural networks in an everyday predictive example. For example, we want to model a neural network for the banking system that predicts debtor risk. For such a problem, we have to build a recurrent neural network that can model patterns over time. RNN will require colossal memory and a large quantity of input data. The neural system will take data sets of previous debtors.

Input variables can be age, income, current debt, etc. and provide the risk factor for the debtor. Each time we ask our neural network for an answer, we also save a set of our intermediate calculations and use them the next time as part of our input. That way, our model will adjust its predictions based on the data that it has seen recently.

PREDICTIVE ANALYTICS



What are the key differences between predictive analytics and machine learning?

As noted, predictive analytics uses advanced mathematics to examine patterns in current and past data in order to predict the future.

Machine learning is a tool that automates predictive modeling by generating training algorithms to look for patterns and behaviors in data without explicitly being told what to look for.

Here are some key differences:

- ML is trained via [supervised and unsupervised learning](#) and is the foundation for advanced technologies such as deep learning and autonomous vehicles.
- Predictive analytics builds on descriptive analytics and diagnostic analytics and is a steppingstone to [prescriptive analytics](#).
- Machine learning algorithms are designed to evolve and improve as they process more data, without being programmed to do so.
- In predictive analytics, data scientists sometimes run the model manually.
- ML works best when given very large data sets. Once a [machine learning algorithm](#) is trained on clean, high-quality data, it can be applied to so-called messy data.
- Predictive analytics depends upon having data that is accurate and complete to build models.

Benefits and challenges of using predictive analytics and machine learning for businesses

Machine learning algorithms can produce more accurate predictions, create cleaner data and empower predictive analytics to work faster and provide more insight with less oversight. Having a strong predictive analysis model and clean data fuels the machine learning application.

While a combination of predictive analytics and ML does not necessarily provide more applications, it does mean that the application can be trusted more. Splitting hairs between the

PREDICTIVE ANALYTICS

two shows that these terms are actually hierarchical and that when combined, they complete one another to strengthen the enterprise.

Challenges: While the techniques associated with both predictive analytics and ML are becoming embedded in software and result in so-called "one-click" forecasting, enterprises will face the usual challenges associated with getting value out of data, starting with the data. Corporate data is error-prone, inconsistent and incomplete. Finding the right data and preparing it for processing is time consuming. Expertise in deploying and interpreting the predictive models is scarce. Moreover, predictive analytics software is expensive, and so is the processing required to make effective models. Lastly, machine learning technologies are evolving at a rapid pace, requiring continuous scrutiny on how and when to upgrade to newer approaches.

What is Random Forest Algorithm?

Random Forest is a famous machine learning algorithm that uses supervised learning methods. You can apply it to both classification and regression problems. It is based on ensemble learning, which integrates multiple classifiers to solve a complex issue and increases the model's performance.

In layman's terms, Random Forest is a classifier that contains several decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset. Instead of relying on a single decision tree, the random forest collects the result from each tree and expects the final output based on the majority votes of predictions.

Working of Random Forest Algorithm

The Working of the Random Forest Algorithm is quite intuitive. It is implemented in two phases: The first is to combine N decision trees with building the random forest, and the second is to make predictions for each tree created in the first phase.

The following steps can be used to demonstrate the working process:

Step 1: Pick M data points at random from the training set.

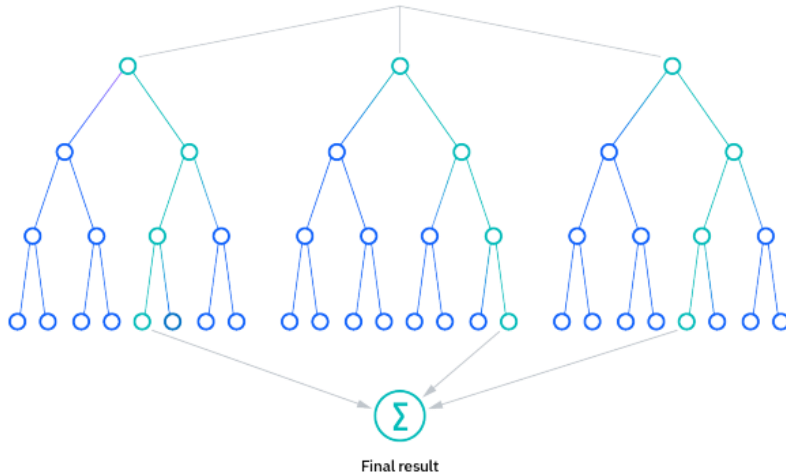
Step 2: Create decision trees for your chosen data points (Subsets).

Step 3: Each decision tree will produce a result. Analyze it.

Step 4: For classification and regression, accordingly, the final output is based on Majority Voting or Averaging, accordingly.

The flowchart below will help you understand better:

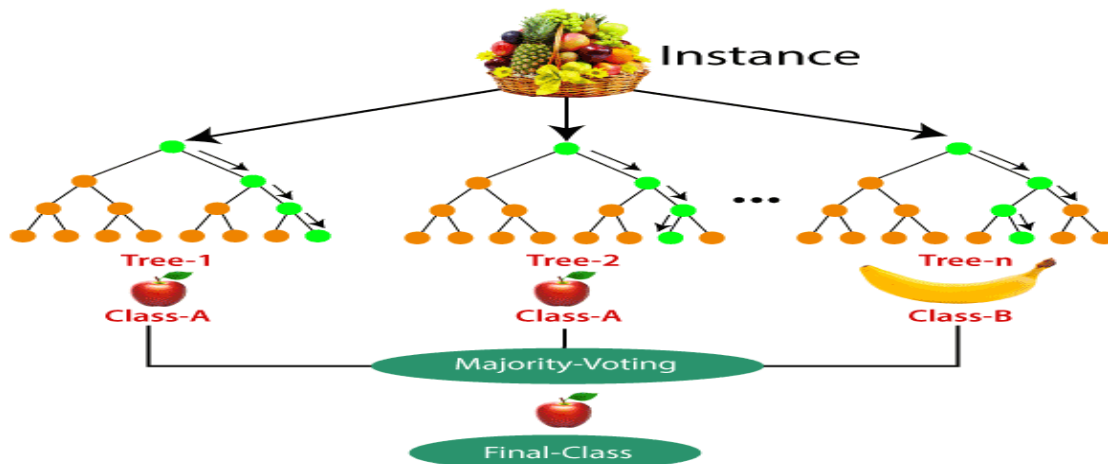
PREDICTIVE ANALYTICS



Example - Consider the following scenario: a dataset containing several fruits images. And the Random Forest Classifier is given this dataset. Each decision tree is given a subset of the dataset to work with. During the training phase, each decision tree generates a prediction result.

The Random Forest classifier predicts the final decision based on most outcomes when a new data point appears.

Consider the following illustration:



How Random Forest Classifier is different from decision trees

Although a random forest is a collection of decision trees, its behavior differs significantly.

PREDICTIVE ANALYTICS

We will differentiate Random Forest from Decision Trees based on 3 Important parameters: Overfitting, Speed, and Process.

1. **Overfitting** - Overfitting is not there as in Decision trees since random forests are formed from subsets of data, and the final output is based on average or majority rating.
2. **Speed** - Random Forest Algorithm is relatively slower than Decision Trees.
3. **Process** - Random forest collects data at random, forms a decision tree, and averages the results. It does not rely on any formulas as in Decision trees.

Ensemble Learning

The ensemble methods in machine learning combine the insights obtained from multiple learning models to facilitate accurate and improved decisions.

Example 1: If you are planning to buy an air-conditioner, would you enter a showroom and buy the air-conditioner that the salesperson shows you? The answer is probably no. In this day and age, you are likely to ask your friends, family, and colleagues for an opinion, do research on various portals about different models, and visit a few review sites before making a purchase decision. In a nutshell, you would not come to a conclusion directly. Instead, you would try to make a more informed decision after considering diverse opinions and reviews. In the case of ensemble learning, the same principle applies.

In learning models, noise, variance, and bias are the major sources of error. The ensemble methods in machine learning help minimize these error-causing factors, thereby ensuring the accuracy and stability of machine learning (ML) algorithms.

Example 2: Assume that you are developing an app for the travel industry. It is obvious that before making the app public, you will want to get crucial feedback on bugs and potential loopholes that are affecting the user experience. What are your available options for obtaining critical feedback? 1) Soliciting opinions from your parents, spouse, or close friends. 2) Asking your co-workers who travel regularly and then evaluating their response. 3) Rolling out your travel and tourism app in beta to gather feedback from non-biased audiences and the travel community.

Taking into account different views and ideas from a wide range of people to fix issues that are limiting the user experience. The ensemble neural network and ensemble algorithm do precisely the same thing.

Ex: Imagine a group of blindfolded people playing the touch-and-tell game, where they are asked to touch and explore a mini donut factory that no one of them has ever seen before. Since they are blindfolded, their version of what a mini donut factory looks like will vary, depending on the parts of the appliance they touch. Now, suppose they are personally asked to describe what they touched. In that case, their individual experiences will give a precise description of specific parts

PREDICTIVE ANALYTICS

of the mini donut factory. Still, collectively, their combined experiences will provide a highly detailed account of the entire equipment.

Ensemble methods in machine learning employ a set of models and take advantage of the blended output, which, compared to a solitary model, will most certainly be a superior option when it comes to prediction accuracy.

Simple Ensemble Methods

Mode: In statistical terminology, "mode" is the number or value that most often appears in a dataset of numbers or values. In this ensemble technique, machine learning professionals use a number of models for making predictions about each data point. The predictions made by different models are taken as separate votes. Subsequently, the prediction made by most models is treated as the ultimate prediction.

The Mean/Average: In the mean/average ensemble technique, data analysts take the average predictions made by all models into account when making the ultimate prediction.

Let's take, for instance, one hundred people rated the beta release of your travel and tourism app on a scale of 1 to 5, where 15 people gave a rating of 1, 28 people gave a rating of 2, 37 people gave a rating of 3, 12 people gave a rating of 4, and 8 people gave a rating of 5.

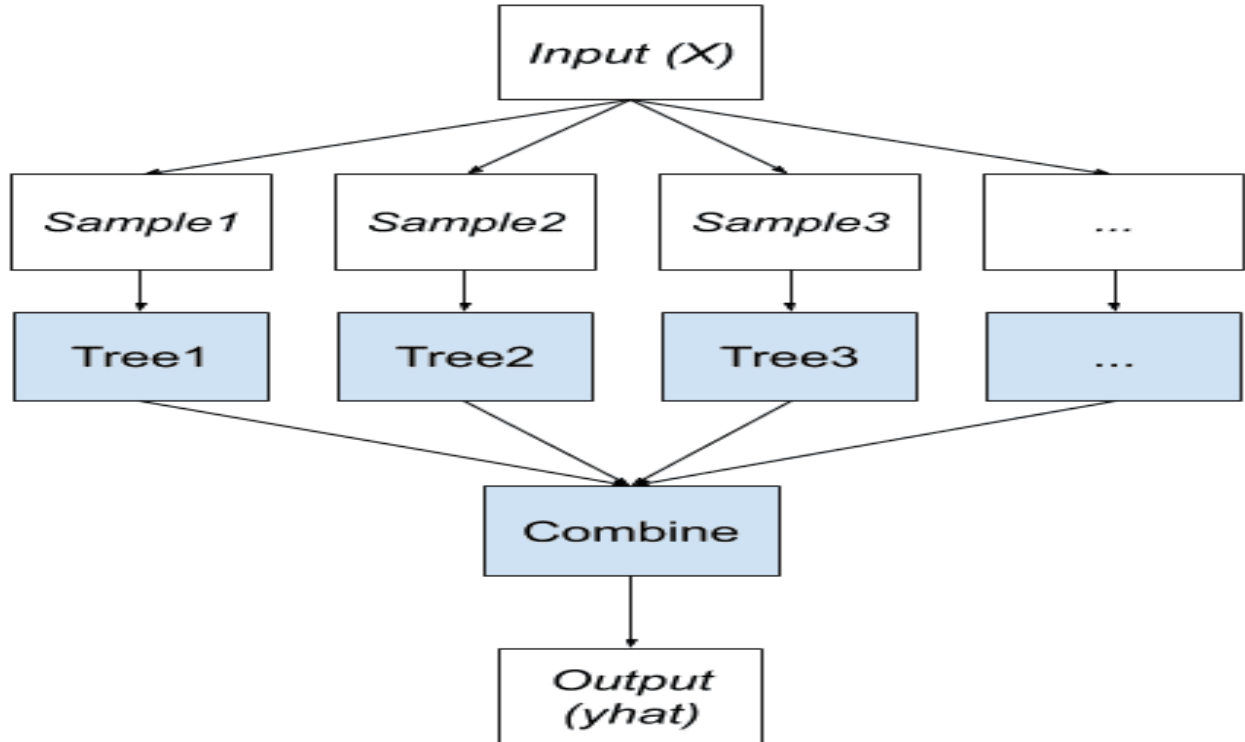
The average in this case is - $(1 * 15) + (2 * 28) + (3 * 37) + (4 * 12) + (5 * 8) / 100 = 2.7$

The Weighted Average: In the weighted average ensemble method, data scientists assign different weights to all the models in order to make a prediction, where the assigned weight defines the relevance of each model. As an example, let's assume that out of 100 people who gave feedback for your travel app, 70 are professional app developers, while the other 30 have no experience in app development. In this scenario, the weighted average ensemble technique will give more weight to the feedback of app developers compared to others.

The three main classes of ensemble learning methods are **bagging**, **stacking**, and **boosting**, and it is important to both have a detailed understanding of each method and to consider them on your predictive modeling project.

- Bagging involves fitting many decision trees on different samples of the same dataset and averaging the predictions.
- Stacking involves fitting many different models types on the same data and using another model to learn how to best combine the predictions.
- Boosting involves adding ensemble members sequentially that correct the predictions made by prior models and outputs a weighted average of the predictions.

Bagging Ensemble



We can summarize the key elements of bagging as follows:

- Bootstrap samples of the training dataset.
- Unpruned decision trees fit on each sample.
- Simple voting or averaging of predictions.

The name Bagging came from the abbreviation of Bootstrap AGGregatING. As the name implies, the two key ingredients of Bagging are bootstrap and aggregation.

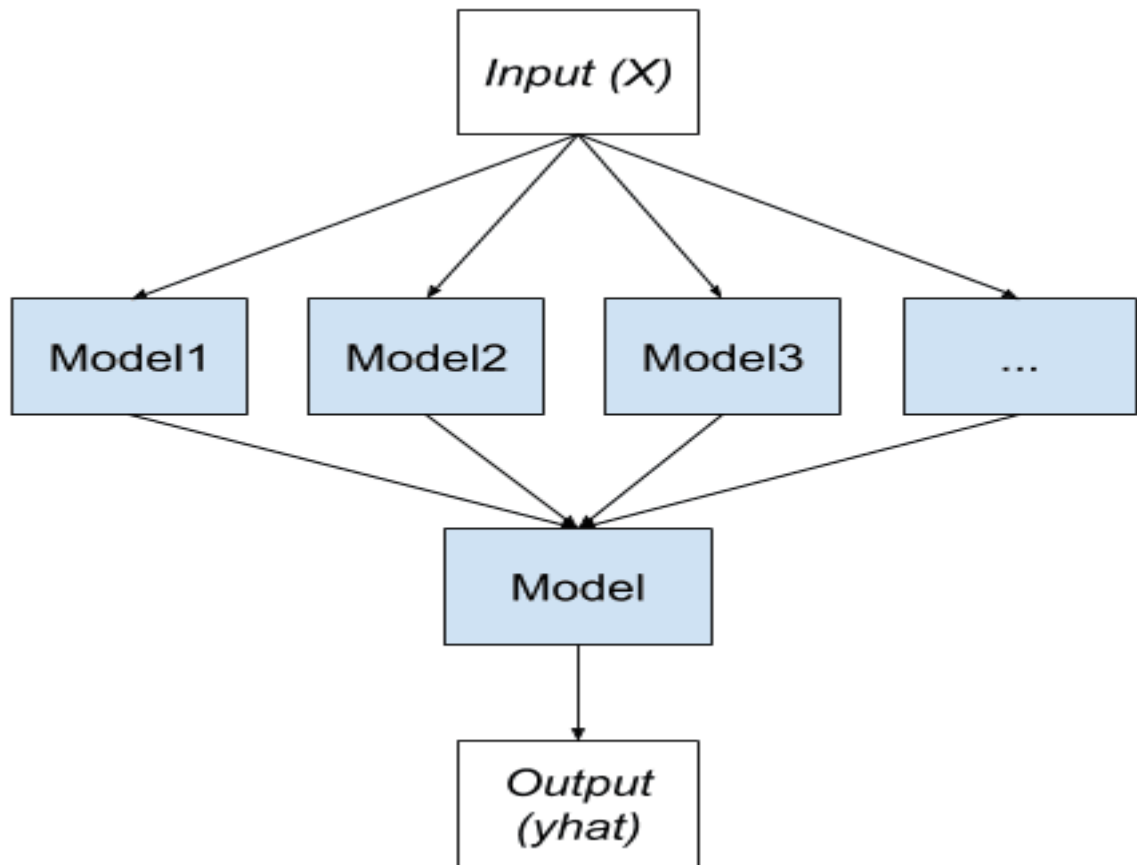
Stacking Ensemble Learning

[Stacked Generalization](#), or stacking for short, is an ensemble method that seeks a diverse group of members by varying the model types fit on the training data and using a model to combine predictions.

We can summarize the key elements of stacking as follows:

- Unchanged training dataset.
- Different machine learning algorithms for each ensemble member.
- Machine learning model to learn how to best combine predictions.

Stacking Ensemble

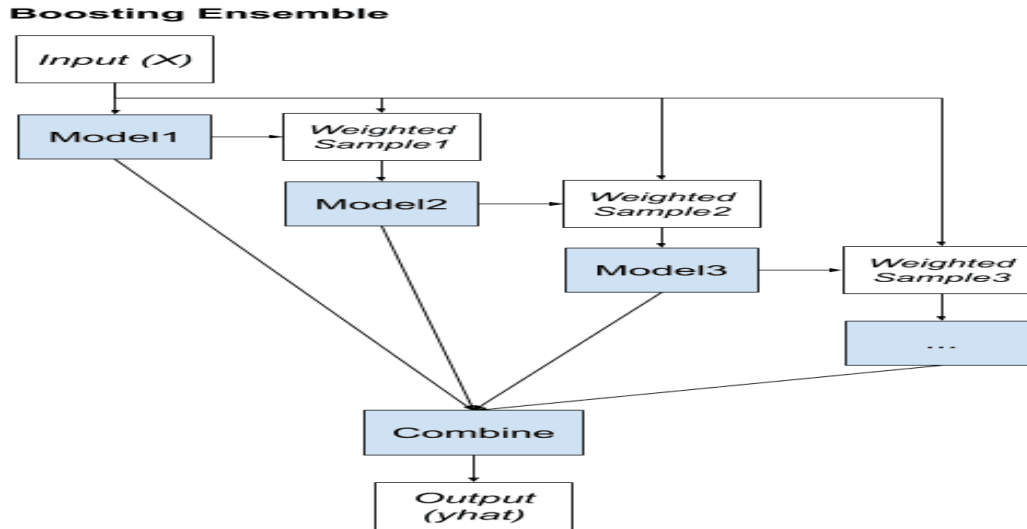


Boosting Ensemble Learning

[Boosting](#) is an ensemble method that seeks to change the training data to focus attention on examples that previous fit models on the training dataset have gotten wrong.

We can summarize the key elements of boosting as follows:

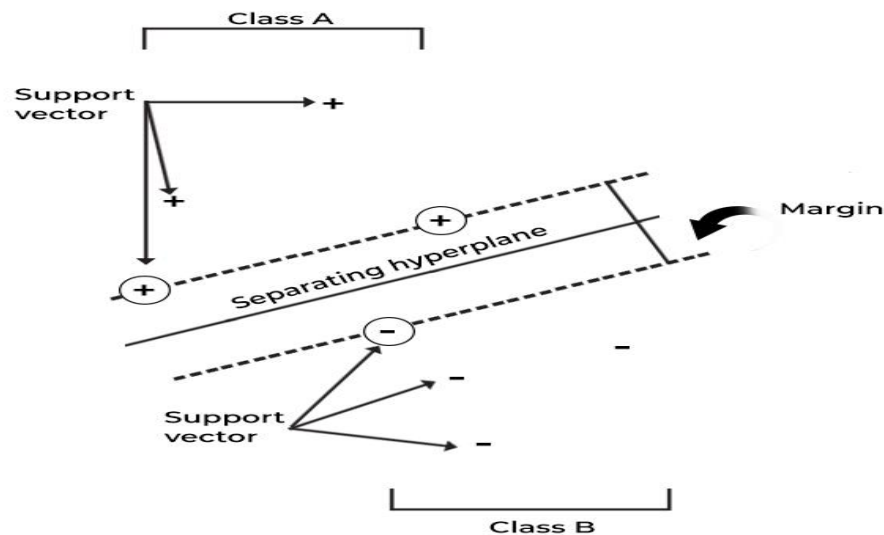
- Bias training data toward those examples that are hard to predict.
- Iteratively add ensemble members to correct predictions of prior models.
- Combine predictions using a weighted average of models.



A support vector machine (SVM) is a machine learning algorithm that uses supervised learning models to solve complex classification, regression, and outlier detection problems by performing optimal data transformations that determine boundaries between data points based on predefined classes, labels, or outputs. SVMs are widely adopted across disciplines such as healthcare, natural language processing, signal processing applications, and speech & image recognition fields.

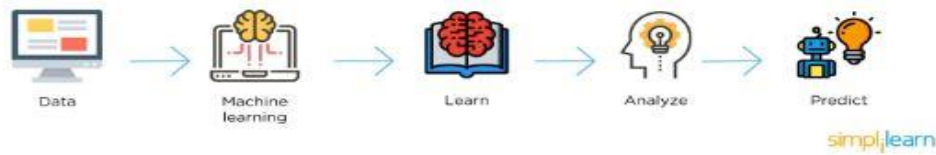


SVMS OPTIMIZE MARGIN BETWEEN SUPPORT VECTORS OR CLASSES



PREDICTIVE ANALYTICS

A computer's ability to learn from data without explicit programming is called [machine learning](#).

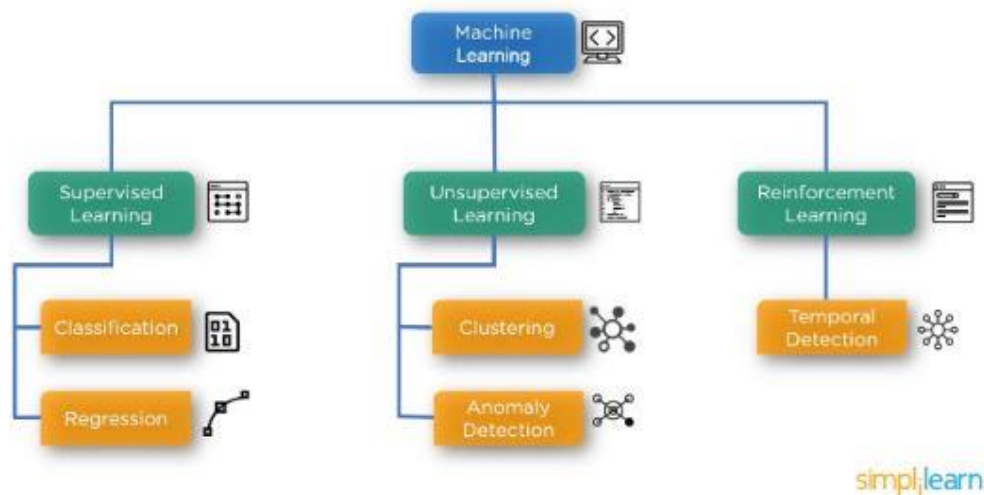


There are three main categories of machine learning algorithms:

- Supervised learning algorithms
- Unsupervised learning algorithms
- Reinforcement learning

Supervised Learning

Supervised learning refers to a data set with known outcomes. If it is unsupervised, there are no known outcomes and you won't have the categories or classes necessary for the machine to learn.



There are two major types of machine learning algorithms in the supervised learning category:

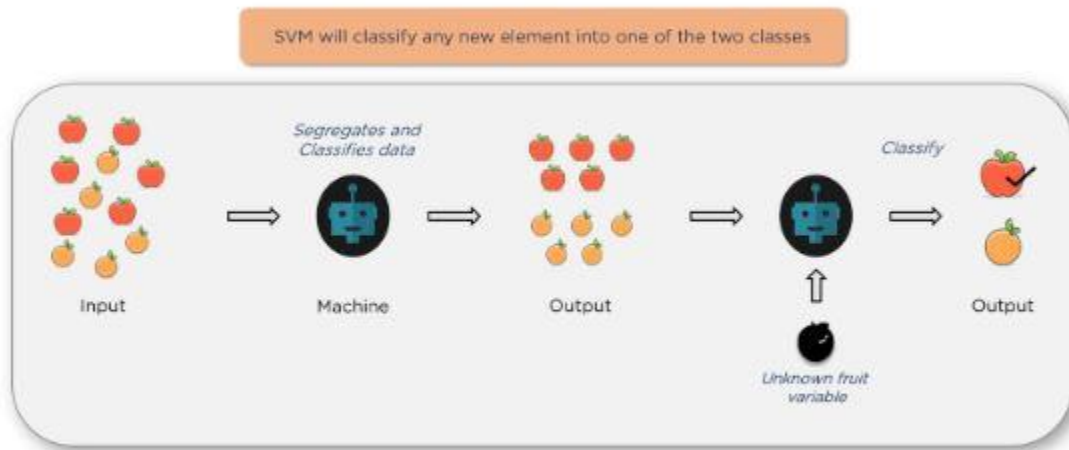
- Classification (which is covered under SVM)
- Regression

Classification Algorithms



What is SVM?

SVM is a type of classification algorithm that classifies data based on its features. An SVM will classify any new element into one of the two classes.



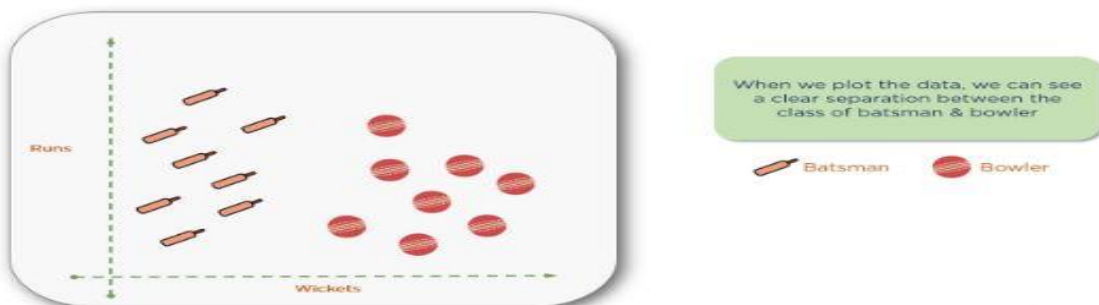
simplilearn

Once you give it some inputs, the algorithm will segregate and classify the data and then create the outputs. When you ingest more new data (an unknown fruit variable in this example), the algorithm will correctly classify the fruit: e.g., “apple” versus “orange”.

Example 1: Linear SVM classification problem with a 2D data set

The goal of this example is to classify cricket players into batsmen or bowlers using the runs-to-wicket ratio. A player with more runs would be considered a batsman and a player with more wickets would be considered a bowler.

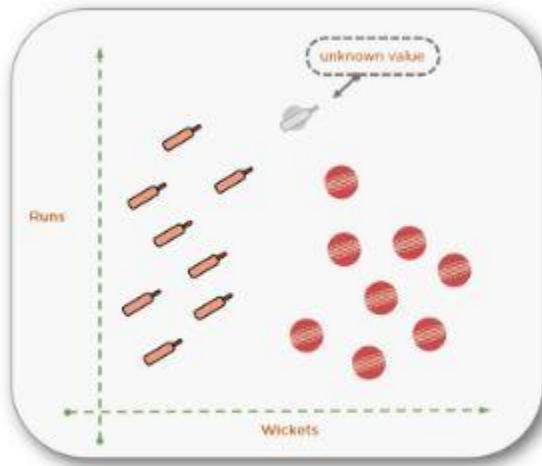
If you take a data set of cricket players with runs and wickets in columns next to their names, you could create a two-dimensional plot showing a clear separation between bowlers and batsmen. Here we present a data set with clear segregation between bowlers versus batsmen to help understand SVM.



simplilearn

PREDICTIVE ANALYTICS

Before separating anything using high-level mathematics, let's look at an unknown value, which is new data being introduced into the dataset without a predesignated classification.

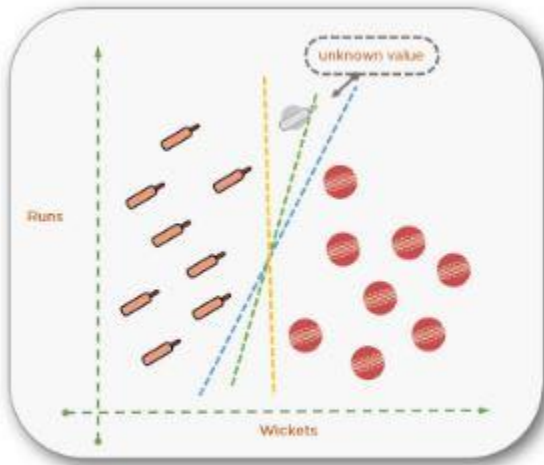


Now, we want to classify a new player variable as a batsman or a bowler

A decision boundary is required in order to classify the new unknown variable

simpl|learn

You can actually draw several boundaries, as shown above. Then, you need to find the line of best fit that clearly separates those two groups. The correct line will help you classify the new data point.

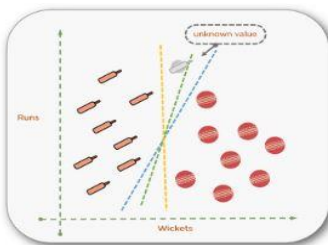


We cannot classify the unknown player into its correct class using multiple separating lines

We need one line that **BEST** separates the data

simpl|learn

You can find the best line by computing the maximum margin from equidistant support vectors. Support vectors in this context simply mean the two points — one from each class that are closest together, but that maximize the distance between them or the margin.



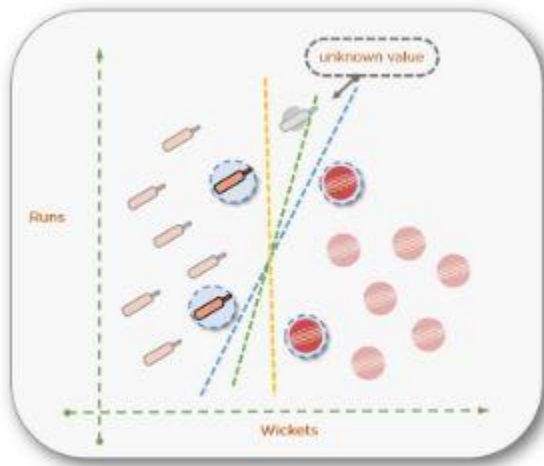
The best line is selected by computing the maximum margin from equidistant Support Vectors

But, what exactly are Support Vectors here?

simpl|learn

PREDICTIVE ANALYTICS

There are a couple of points at the top that are pretty close to one another, and similarly at the bottom of the graph. Shown below are the points that you need to consider. The rest of the points are too far away. The bowler points to the right and the batsman points to the left.

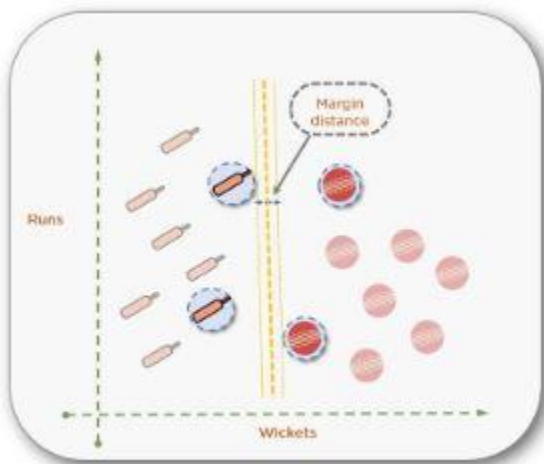


Support vectors are the points which are very close to a dividing line

Using the support vectors we can select the best line to divide the data

simpl|learn

Mathematically, you can calculate the distance among all of these points and minimize that distance. Once you pick the support vectors, draw a dividing line, and then measure the distance from each support vector to the line. The best line will always have the greatest margin or distance between the support vectors.

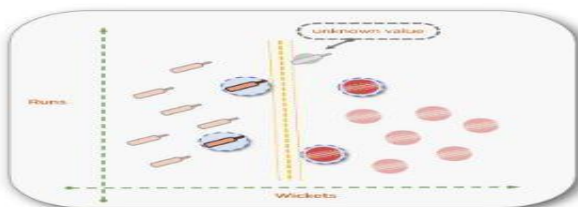


Margin is the distance between the support vectors and a dividing line

The best line will always have the greatest margin distance between the support vectors

simpl|learn

For instance, if you consider the yellow line as a decision boundary, the player with the new data point is the bowler. But, as the margins don't appear to be maximum, you can come up with a better line.



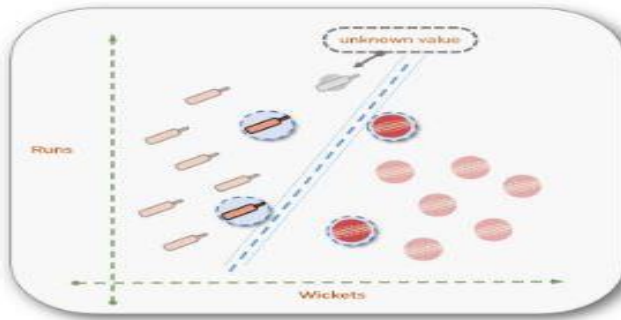
If we consider the yellow line as a decision boundary, the player will be classified as a bowler

The margins don't appear to be at maximum with this line, so let's consider the blue line

simpl|learn

PREDICTIVE ANALYTICS

Use other support vectors, draw the decision boundary between those, and then calculate the margin. Notice now that the unknown data point would be considered a batsman.

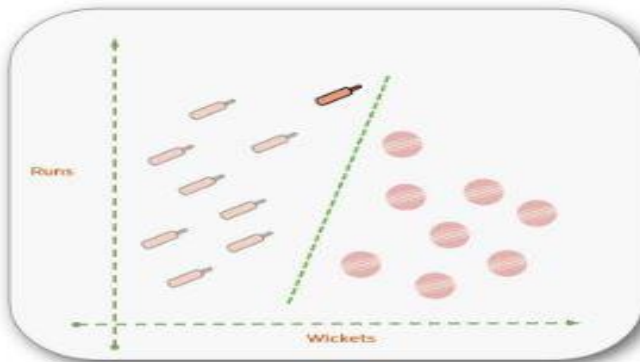


The blue line here will classify the player as a batsman. Also, the margin distance appears to be the same as the yellow line.

Let us now consider the green line and compare its margin distance

simplilearn

If you look at the green decision boundary, the line appears to have a maximum margin compared to the other two. This is the boundary of greatest margin and when you classify your unknown data value, you can see that it clearly belongs to the batsman's class. The green line divides the data perfectly because it has the maximum margin between the support vectors. At this point, you can be confident with the classification — the new data point is indeed a batsman.



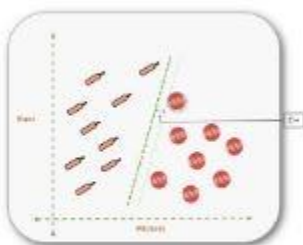
The green line divides the data perfectly because the margin between Support Vectors are maximum

Using the green line as the decision boundary, the player is classified as a batsman

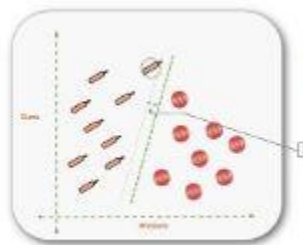


simplilearn

The hyperplane with the maximum distance from the support vectors is the one you want. Sometimes called the positive hyperplane (D^+), it is the shortest distance to the closest positive point and (D^-), or the negative hyperplane, which is the shortest distance to the closest negative point.



D^+ is the shortest distance to the closest positive point

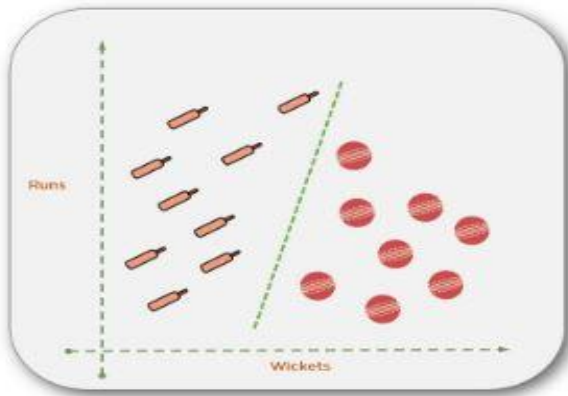


D^- is the shortest distance to the closest negative point

simplilearn

This problem set is two-dimensional because the classification is only between two classes. It is called a linear SVM.

PREDICTIVE ANALYTICS



This problem set is 2 Dimensional because classification is only between 2 classes

2 Dimensional applications of SVM are called *linear SVM*

simplilearn

The following are the steps to make the classification:



simplilearn

Applications of Support Vector Machine



Linear SVM vs Non-Linear SVM

Linear SVM

The data points are separated using a single line

Data is classified with the help of a hyperplane.

difficult to classify more than two labels.

Non Linear SVM

The data points are hard to separate, so other shapes are used as a decision boundary.

Kernels are used to classify data points.

Used for classifying more than two labels.

PREDICTIVE ANALYTICS

DATA MINING

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called *Knowledge Discovery in Database (KDD)*. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.



Types of Data Mining

Data mining can be performed on the following types of data:

Relational Database

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data search ability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical

purposes and helps in decision-making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc. One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.

PREDICTIVE ANALYTICS

- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

What is Data Interpretation

Data interpretation refers to the process of taking raw data and transforming it into useful information. This involves analyzing the data to identify patterns, trends, and relationships, and then presenting the results in a meaningful way. Data interpretation is an essential part of data analysis, and it is used in a wide range of fields, including business, marketing, healthcare, and many more.

Importance of Data Interpretation in Today's World

Data interpretation is critical to making informed decisions and driving growth in today's data-driven world. With the increasing availability of data, companies can now gain valuable insights into their operations, customer behavior, and market trends. Data interpretation allows businesses to make informed decisions, identify new opportunities, and improve overall efficiency.

Importance of Data Interpretation in Today's World

Data interpretation is critical to making informed decisions and driving growth in today's data-driven world. With the increasing availability of data, companies can now gain valuable insights into their operations, customer behavior, and market trends. Data interpretation allows businesses to make informed decisions, identify new opportunities, and improve overall efficiency.

Methods of Data Interpretation

There are several data interpretation methods, including descriptive statistics, inferential statistics, and visualization techniques.

Descriptive Statistics

Descriptive statistics involve summarizing and presenting data in a way that makes it easy to understand. This can include calculating measures such as mean, median, mode, and standard deviation.

PREDICTIVE ANALYTICS

Inferential Statistics

Inferential statistics involves making inferences and predictions about a population based on a sample of data. This type of data interpretation involves the use of statistical models and algorithms to identify patterns and relationships in the data.

Visualization Techniques

Visualization techniques involve creating visual representations of data, such as graphs, charts, and maps. These techniques are particularly useful for communicating complex data in an easy-to-understand manner and identifying data patterns and trends.

Benefits of Data Interpretation

Data interpretation plays a crucial role in decision-making and helps organizations make informed choices. There are numerous benefits of data interpretation, including:

- **Improved decision-making:** Data interpretation provides organizations with the information they need to make informed decisions. By analyzing data, organizations can identify trends, patterns, and relationships that they may not have been able to see otherwise.
- **Increased efficiency:** By automating the data interpretation process, organizations can save time and improve their overall efficiency. With the right tools and methods, data interpretation can be completed quickly and accurately, providing organizations with the information they need to make decisions more efficiently.
- **Better collaboration:** Data interpretation can help organizations work more effectively with others, such as stakeholders, partners, and clients. By providing a common understanding of the data and its implications, organizations can collaborate more effectively and make better decisions.
- **Increased accuracy:** Data interpretation helps to ensure that data is accurate and consistent, reducing the risk of errors and miscommunication. By using data interpretation techniques, organizations can identify errors and inconsistencies in their data, making it possible to correct them and ensure the accuracy of their information.
- **Enhanced transparency:** Data interpretation can also increase transparency, helping organizations demonstrate their commitment to ethical and responsible data management. By providing clear and concise information, organizations can build trust and credibility with their stakeholders.
- **Better resource allocation:** Data interpretation can help organizations make better decisions about resource allocation. By analyzing data, organizations can identify areas where they are spending too much time or money and make adjustments to optimize their resources.
- **Improved planning and forecasting:** Data interpretation can also help organizations plan for the future. By analyzing historical data, organizations can identify trends and patterns that inform their forecasting and planning efforts.

Data Interpretation Process

PREDICTIVE ANALYTICS

Data interpretation is a process that involves several steps, including:

- **Data collection:** The first step in data interpretation is to collect data from various sources, such as surveys, databases, and websites. This data should be relevant to the issue or problem the organization is trying to solve.
- **Data preparation:** Once data is collected, it needs to be prepared for analysis. This may involve cleaning the data to remove errors, missing values, or outliers. It may also include transforming the data into a more suitable format for analysis.
- **Data analysis:** The next step is to analyze the data using various techniques, such as statistical analysis, visualization, and modeling. This analysis should be focused on uncovering trends, patterns, and relationships in the data.
- **Data interpretation:** Once the data has been analyzed, it needs to be interpreted to determine what the results mean. This may involve identifying key insights, drawing conclusions, and making recommendations.
- **Data communication:** The final step in the data interpretation process is to communicate the results and insights to others. This may involve creating visualizations, reports, or presentations to share the results with stakeholders.

Data Interpretation Tools

Data interpretation is a crucial step in the data analysis process, and the right tools can make a significant difference in accuracy and efficiency. Here are a few tools that can help you with data interpretation:

- **Layer:** Layer is a free Google Sheets add-on that equips you with the tools to increase the efficiency and data quality of your processes on top of Google Sheets.
 - Share parts of your spreadsheet, including sheets or even cell ranges, with different collaborators or stakeholders.
 - Review and approve edits by collaborators to their respective sheets before merging them back with your master spreadsheet.
 - Integrate popular tools and connect your tech stack to sync data from different sources, giving you a timely, holistic view of your data.
- **Google Sheets:** Google Sheets is a free, web-based spreadsheet application that allows users to create, edit, and format spreadsheets. It provides a range of features for data interpretation, including functions, charts, and pivot tables.
- **Microsoft Excel:** Microsoft Excel is a spreadsheet software widely used for data interpretation. It provides various functions and features to help you analyze and interpret data, including sorting, filtering, pivot tables, and charts.
- **Tableau:** Tableau is a data visualization tool that helps you see and understand your data. It allows you to connect to various data sources and create interactive dashboards and visualizations to communicate insights.
- **Power BI:** Power BI is a business analytics service that provides interactive visualizations and business intelligence capabilities with an easy interface for end users to create their own reports and dashboards.
- **R:** R is a programming language and software environment for statistical computing and graphics. It is widely used by statisticians, data scientists, and researchers to analyze and interpret data.

PREDICTIVE ANALYTICS

Data Reduction: The method of data reduction may achieve a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

INTRODUCTION:

Data reduction is a technique used in data mining to reduce the size of a dataset while still preserving the most important information. This can be beneficial in situations where the dataset is too large to be processed efficiently, or where the dataset contains a large amount of irrelevant or redundant information.

There are several different data reduction techniques that can be used in data mining, including:

1. **Data Sampling:** This technique involves selecting a subset of the data to work with, rather than using the entire dataset. This can be useful for reducing the size of a dataset while still preserving the overall trends and patterns in the data.
2. **Dimensionality Reduction:** This technique involves reducing the number of features in the dataset, either by removing features that are not relevant or by combining multiple features into a single feature.
3. **Data Compression:** This technique involves using techniques such as lossy or lossless compression to reduce the size of a dataset.
4. **Data Discretization:** This technique involves converting continuous data into discrete data by partitioning the range of possible values into intervals or bins.
5. **Feature Selection:** This technique involves selecting a subset of features from the dataset that are most relevant to the task at hand.
6. It's important to note that data reduction can have a trade-off between the accuracy and the size of the data. The more data is reduced, the less accurate the model will be and the less generalizable it will be.

In conclusion, data reduction is an important step in data mining, as it can help to improve the efficiency and performance of machine learning algorithms by reducing the size of the dataset. However, it is important to be aware of the trade-off between the size and accuracy of the data, and carefully assess the risks and benefits before implementing it.

What is classification

The term "classification" is usually used when there are exactly two target classes called binary classification. When more than two classes may be predicted, specifically in pattern recognition problems, this is often referred to as multinomial classification. However, multinomial classification is also used for categorical response data, where one wants to predict which category amongst several categories has the instances with the highest probability.

Classification is one of the most important tasks in data mining. It refers to a process of assigning pre-defined class labels to instances based on their attributes. There is a similarity between classification and clustering, it looks similar, but it is different. The major difference between

classification and clustering is that classification includes the leveling of items according to their membership in pre-defined groups. Let's understand this concept with the help of an example; suppose you are using a self-organizing map neural network algorithm for image recognition where there are 10 different kinds of objects. If you label each image with one of these 10 classes, the classification task is solved.

On the other hand, clustering does not involve any labeling. Assume that you are given an image database of 10 objects and no class labels. Using a clustering algorithm to find groups of similar-looking images will result in determining clusters without object labels.

Classification of data mining

These are given some of the important data mining classification methods:

Logistic Regression Method

The logistic Regression Method is used to predict the response variable.

K-Nearest Neighbors Method

K-Nearest Neighbors Method is used to classify the datasets into what is known as a K observation. It is used to determine the similarities between the neighbours.

Naive Bayes Method

The Naive Bayes method is used to scan the set of data and locate the records wherein the predictor values are equal.

Neural Networks Method

The Neural Networks resemble the structure of our brain called the Neuron. The sets of data pass through these networks and finally come out as output. This neural network method compares the different classifications. Errors that occur in the classifications are further rectified and are fed into the networks. This is a recurring process.

Discriminant Analysis Method

In this method, a linear function is built and used to predict the class of variables from observation with the unknown class.

PREDICTIVE ANALYTICS

What is clustering

Clustering refers to a technique of grouping objects so that objects with the same functionalities come together and objects with different functionalities go apart. In other words, we can say that clustering is a process of portioning a data set into a set of meaningful subclasses, known as clusters. Clustering is the same as classification in which data is grouped. Though, unlike classification, the groups are not previously defined. Instead, the grouping is achieved by determining similarities between data according to characteristics found in the real data. The groups are called Clusters.

Methods of clustering

- Partitioning methods
- Hierarchical clustering
- Fuzzy Clustering
- Density-based clustering
- Model-based clustering

Difference between Classification and Clustering

Classification	Clustering
Classification is a supervised learning approach where a specific label is provided to the machine to classify new observations. Here the machine needs proper testing and training for the label verification.	Clustering is an unsupervised learning approach where grouping is done on similarities basis.
Supervised learning approach.	Unsupervised learning approach.
It uses a training dataset.	It does not use a training dataset.
It uses algorithms to categorize the new data as per the observations of the training set.	It uses statistical concepts in which the data set is divided into subsets with the same features.
In classification, there are labels for training data.	In clustering, there are no labels for training data.

PREDICTIVE ANALYTICS

Its objective is to find which class a new object belongs to form the set of predefined classes.	Its objective is to group a set of objects to find whether there is any relationship between them.
It is more complex as compared to clustering.	It is less complex as compared to clustering.

Association rule mining is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

Association rule mining is a type of unsupervised machine learning that discovers interesting relationships between variables in large datasets. It is a rule-based approach that finds association rules, which are if-then statements that describe the relationship between two or more items.

Types Of Association Rules In Data Mining

There are typically four different types of association rules in data mining. They are

- Multi-relational association rules
- Generalized Association rule
- Interval Information Association Rules
- Quantitative Association Rules

Multi-Relational Association Rule

Also known as MRAR, multi-relational association rule is defined as a new class of association rules that are usually derived from different or multi-relational databases. Each rule under this class has one entity with different relationships that represent the indirect relationships between entities.

Generalized Association Rule

Moving on to the next type of association rule, the generalized association rule is largely used for getting a rough idea about the interesting patterns that often tend to stay hidden in data.

Quantitative Association Rules

This particular type is actually one of the most unique kinds of all the four association rules available. What sets it apart from the others is the presence of numeric attributes in at least one attribute of quantitative association rules. This is in contrast to the generalized association rule, where the left and right sides consist of categorical attributes.

Algorithms Of Associate Rule In Data Mining

There are mainly three different types of algorithms that can be used to generate associate rules in data mining. Let's take a look at them.

- Apriori Algorithm

Apriori algorithm identifies the frequent individual items in a given database and then expands them to larger item sets, keeping in check that the item sets appear sufficiently often in the database.

- Eclat Algorithm

ECLAT algorithm is also known as Equivalence Class Clustering and bottomup. Lattice Traversal is another widely used method for associate rule in data mining. Some even consider it to be a better and more efficient version of the Apriori algorithm.

Cause and Effect Relationship:

A cause and effect modeling is done to uncover patterns in data of the business organizations.

There are many different types of causal patterns in the world. Below are six patterns that are embedded in many concepts. Causality in the real world seldom falls into one neat pattern or another. The patterns often work together or different parts of a system entail different patterns—making the causality even more complex!

- Linear Causality – Cause precedes effect; sequential pattern. Direct link between cause and effect. Has a clear beginning and a clear ending. Effect can be traced back to one cause. One cause and one effect; additional causes or effects turn this pattern into domino causality
- Domino Causality – Sequential unfolding of effects over time. An extended linear pattern that results in direct and indirect effects. Typically has a clear beginning and a clear ending. Can be branching where there is more than one effect of a cause (and these may go on to have multiple effects and so on.). Branching forms can be traced back to “stem” causes. Anticipating outcomes involves deciding how far to trace effects. Short-sightedness can lead to unintended effects.
- Cyclic Causality – One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but eventually impacts the first thing). Involves a repeating pattern. Involves feedback loops. May be sequential or may be simultaneous. Typically no clear beginning or ending (Sometimes you can look back in time to a beginning but often that results in the classic ‘which came first, the chicken or the egg’ problem.).
- Spiraling Causality – One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but

PREDICTIVE ANALYTICS

eventually impacts the first thing) with amplification or de-amplification of effects. Involves feedback loops. It is sequential as each event is a reaction to the one before it. Often a clear beginning and ending. It is difficult to anticipate outcomes of later feedback loops during earlier feedback loops.

- Relational Causality – Two things work in relation to each other to cause an outcome. It often involves two variables in comparison to each other. There may be a relationship of balance, equivalence, similarity or there may be a relationship of difference. If one thing changes, so does the relationship, therefore so does the outcome. If two things change but keep the same relationship, the outcome doesn't change.
- Mutual Causality – Two things impact each other. The impact can be positive for both, negative for both, or positive for one and negative for the other. The causes and effects are often simultaneous, but can be sequential. May be event-based or may be a relationship over time (such as the moss and the algae in lichen).



A cause and effect analysis is an attempt to understand why things happen as they do. People in many professions—accident investigators, scientists, historians, doctors, newspaper reporters, automobile mechanics, educators, police detectives—spend considerable effort trying to understand the causes and effects of human behavior and natural phenomena to gain better control over events and over ourselves. If we understand the causes of accidents, wars, and natural disasters, perhaps we can avoid them in the future. If we understand the consequences of our own behavior, perhaps we can modify our behavior in a way that will allow us to lead happier, safer lives.

How to Use Cause and Effect Analysis to Solve Business Problems

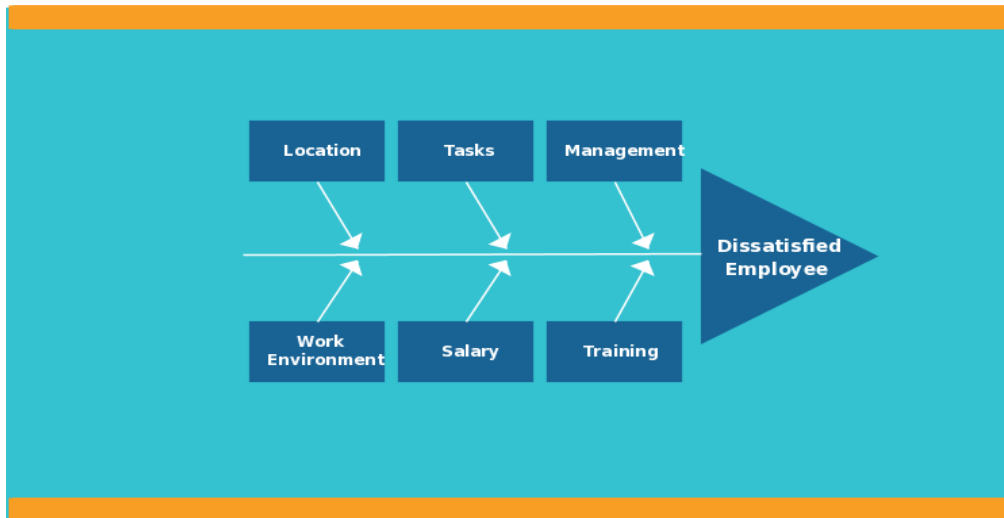
A problem in your business can be an opportunity for the business to grow or can be the setback that leads to failure. It all depends on how you embark on problem-solving.

Step 1: Problem Identification

The way to use this tool is very simple. You first need to identify the problem area that needs analysis. A good way to do this is by giving a brief description of the current business situation and the consequences as well as the reasons for why they occurred.

PREDICTIVE ANALYTICS

Step 2: Main Problem Causes



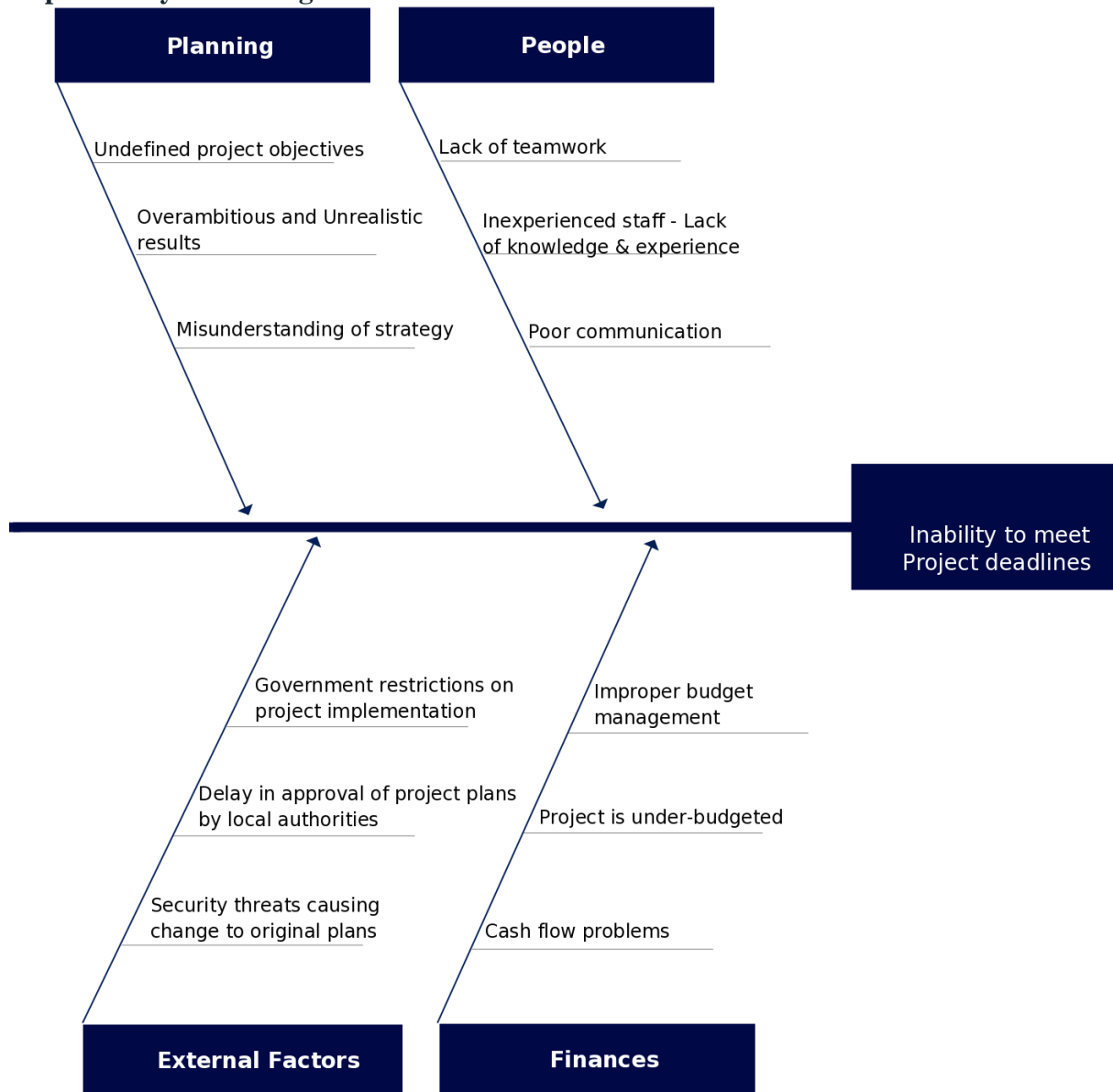
The next step is identifying the main causes of the problem, for example, the people, the procedures your business utilized as well as the materials or equipment. Here, you will need to do a lot of brainstorming to come up as many reasons as you possibly can.

Step3: Identify Plausible Sub-causes of the Main Causes

Thereafter, you can narrow down to further plausible causes of the main problems that you've listed. This is important for the problem-solving stage of analyzing the diagram, as you will know what to do exactly to rectify the situation. This can include issues like wrong norms and values that could be as a result of poor training of employees.

PREDICTIVE ANALYTICS

Step 4: Analyze the Diagram



The cause and effect analysis uses brainstorming and critical analysis by way of [visual representation to enable problem-solving](#).

Data Simulation

Data simulation is the process of taking a large amount of data and using it to mimic real-world scenarios or conditions. In technical terms, it could be described as the generation of random numbers or data from a stochastic process which is stated as a distribution equation (e.g., Normal: $X \sim N(\mu, \sigma^2)$). It can be used to predict future events, determine the best course of action or validate AI/ML models.

Benefits of Data Simulation

Data simulation has proven highly valuable across nearly every industry and field of study, with business executives, engineers and researchers all using it in their work. Among other benefits, data simulation can:

1. Enable the creation of comprehensive models of complex, dynamic systems;
2. Empower data-driven decision making and strategic planning;
3. Help test hypotheses, understand relationships, and improve predictions;
4. Allow the study of phenomena that is difficult or impossible to investigate directly; and
5. Generate synthetic data that is representative of specific populations or conditions which can then be used for ML and AI development.

Use Cases for Data Simulation

With the advent of high-quality synthetic data generation technology and state-of-the-art ML/AI models, there are some fascinating use cases for data simulation that have emerged in recent years. Here are some specific examples across various fields:

Software Development

A key part of developing any software is testing how it will perform under different conditions. By creating data simulations that mimic real-world conditions, developers can put the software through its paces and identify any potential problems. This process can be used to test everything from the user interface to the backend algorithms.

Oil & Gas

Data simulation is increasingly being used in the oil and gas industry, too. By creating models of reservoirs, geologists can better understand how oil and gas flow through rock and whether they're present in different geological strata. These models can be used to predict what will happen when new wells are drilled, and they can help engineers design better production facilities, too.

Companies and researchers also study the impact of environmental factors on the industry. By simulating the effects of climate change, researchers gain better understanding of how rising temperatures might affect the production of oil and gas.

Manufacturing

Data simulation is also being used to create “[digital twins](#)” which are virtual copies of physical objects, such as a car or production factory. These models enable the study of real-world objects and their operations without ever touching them. Manufacturers can easily identify the most efficient and effective production process for a particular product, and avoid disruptions as they transition to new methods.

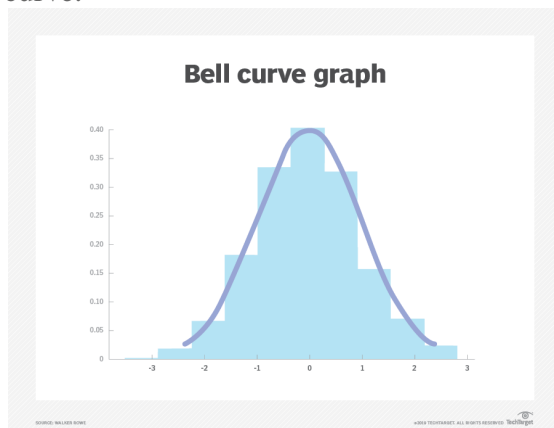
Autonomous Vehicles

And of course, we can't talk about data simulation without acknowledging its most high-profile use case: the training of self-driving cars, drones and robots. Trying to test and train these systems in the real-world is slow, costly and dangerous. But with synthetic data you can create virtual training environments for improving these emerging technologies.

What is a Monte Carlo simulation

A Monte Carlo simulation is a mathematical technique that simulates the range of possible outcomes for an uncertain event. These predictions are based on an estimated range of values instead of a fixed set of values and evolve randomly. Computers use Monte Carlo simulations to analyze data and predict a future outcome based on a course of action.

First, Monte Carlo simulations use a probability distribution for any variable that has inherent uncertainty. Then, it recalculates the results many times, using a different set of random numbers within the estimated range each time. This process generates many probable outcomes, which become more accurate as the number of inputs grows. In other words, the different outcomes form a [normal distribution](#) or [bell curve](#), where the most common outcome is in the middle of the curve.



The 4 steps in a Monte Carlo simulation

Although they might vary from case to case, the general steps to a Monte Carlo simulation are as follows:

1. Build the model. Determine the mathematical model or transfer algorithm.
2. Choose the variables to simulate. Pick the variables, and determine an appropriate probability distribution for each random variable.
3. Run repeated simulations. Run the random variables through the mathematical model to perform many iterations of the simulation.
4. Aggregate the results, and determine the mean, standard deviation and variant to determine if the result is as expected. Visualize the results on a [histogram](#).

PREDICTIVE ANALYTICS

Industry use cases for a Monte Carlo simulation include the following:

- **Finance**, such as risk assessment and long-term forecasting.
- **Project management**, such as estimating the duration or cost of a project.
- **Engineering and physics**, such as analyzing weather patterns, traffic flow or energy distribution.
- **Quality control and testing**, such as estimating the reliability and failure rate of a product.
- **Healthcare and biomedicine**, such as modeling the spread of diseases.

Some of those use cases specific to IT are the following:

- **Network and system design.** Monte Carlo simulations can be used to model different designs, identify potential bottlenecks, and perform capacity planning and resource allocation.
- **Artificial intelligence.** Monte Carlo simulations provide the basis for resampling techniques for estimating the accuracy of a model on a given data set.
- **Cybersecurity.** Monte Carlo simulations can be used to simulate [different cyber attacks](#), evaluate the probability of them occurring, evaluate their hypothetical impact and identify vulnerabilities in IT systems.
- **Performance testing.** Monte Carlo simulations can be used for [load testing applications](#) and estimating the potential impact for increased usage or scaling.

Other advantages of Monte Carlo simulations include the following:

- **Improve decision-making.** Monte Carlo simulations help users make decisions with a degree of confidence.
- **Solve complex problems simply.** Monte Carlo simulations show both what could happen and how likely each outcome is
- **Visualize the range of possible outcomes and their likelihood of occurring.** Monte Carlo simulations make it easy to visualize what the result of a standard decision or outcome might be next to the result of an unusual outcome.

Discriminant event simulation

Discriminant event simulation is a type of simulation technique used in the field of operations research and management science to model and analyze complex systems or processes that involve both deterministic and random events. It is a combination of two techniques: discriminant analysis and event simulation.

Discriminant analysis is a statistical technique used to classify observations into two or more groups based on a set of predictor variables. In discriminant event simulation, the classification is based on a set of decision rules or policies that determine the actions to be taken in response to different events or scenarios.

PREDICTIVE ANALYTICS

Event simulation, on the other hand, is a technique used to model and simulate the behavior of a system over time by representing its components as discrete events or processes. Events are triggered by external or internal factors, and can lead to changes in the state of the system or the occurrence of new events.

The basic steps in a discriminant event simulation are:

Define the problem: Define the system or process to be modeled, and identify the decision rules or policies that will be used to classify observations and determine the actions to be taken.

Define the model: Represent the system as a set of components or processes, and define the events that can occur and the rules that govern their occurrence.

Define the input data: Define the input data required for the simulation, including the probabilities of different events and the values of the predictor variables.

Generate random samples: Generate a large number of random samples from the input data, and simulate the behavior of the system over time for each sample.

Analyze the results: Analyze the results of the simulation to gain insights into the behavior of the system and the performance of the decision rules or policies. This may involve statistical analysis, visualization, or other techniques.

Discriminant event simulation can be used in a wide range of applications, such as finance, marketing, supply chain management, and health care, to model and analyze complex systems that involve both deterministic and random events. For example, it can be used to model customer behavior and evaluate marketing strategies, optimize inventory levels and supply chain operations, or simulate the spread of infectious diseases and evaluate public health policies.

Advantages of discriminant event simulation:

Flexibility: Discriminant event simulation is a flexible tool that can be used to model a wide range of systems or processes, including those with complex or non-linear relationships between variables.

Quantitative analysis: Discriminant event simulation provides a quantitative way to analyze the behavior of a system or process and evaluate the performance of decision rules or policies.

Risk assessment: Discriminant event simulation can be used to assess the risks and uncertainties associated with a system or process, and to identify the factors that contribute most to the variability in the outputs.

PREDICTIVE ANALYTICS

Evaluation of alternatives: Discriminant event simulation can be used to evaluate the performance of different decision rules or policies, and to identify the optimal solutions based on the results of the simulations.

Limitations of discriminant event simulation:

Computationally intensive: Discriminant event simulation can be computationally intensive, especially when simulating complex systems or using a large number of random samples.

Input data requirements: Discriminant event simulation requires accurate and reliable input data to ensure that the results of the simulations are meaningful and relevant.

Simplifications and assumptions: Discriminant event simulation is based on simplifications and assumptions, which may not accurately reflect the complexity of the real system being modeled.