**Statistical Models in Business Analytics**

**What is Statistical Modeling?**

Statistical Modeling is the formalization of relationships between variables in the form of mathematical equations. It is basically about finding out the variable. It describes how one or more variables are related to one or more other variables. Here, the variables are not accurately related but could be stochastically related.

In simpler terms, a variable is nothing but an attribute. An attribute becomes a person’s height, weight and age. Height and age are probabilistic in nature. A 30-year old person has higher chances of being 4 ft tall. Similarly, when you are aware of a person who is 13 years of age, he has higher chances of being 6 ft tall.

The whole purpose of statistical modeling is not about the research, it ultimately comes down to providing an insight to solutions. It involves analyzing the data and applying it in different circumstances.

**What is Regression Modeling?**

Regression modeling is all about sourcing the relationship between two variables. More specifically, regression helps one understand how the value of the dependent variable changes while any one of the independent variable varies, while the other independent variables are held fixed. For instance, time is an independent variable while sales and velocity are dependent on certain factors. Hence, the goal is to find out the relationship between the two.

There are certain equations in the regression model, it being the linear, multivariate and logistic regression. Logistic regression is similar to regression wherein there are two variables, therefore classifying itself as a probabilistic statistical model. It is used in describing the parameters of a qualitative response model.

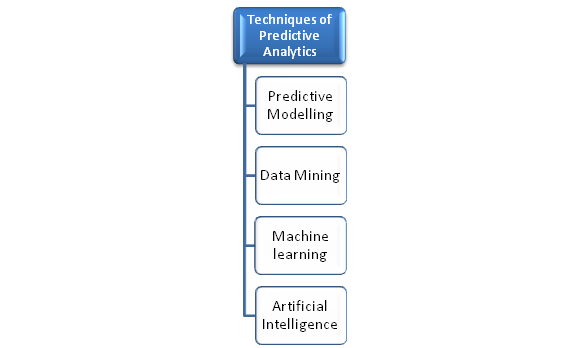
**Understanding Analytics in Business**

The entire operation of analytics boils down to 3 simple models- Predictive, descriptive and decision model. As the name suggests, it enables one to comprehend the future. For instance, system failure, credit worthiness, fraud come under the predictive model which is gaining world wide popularity today. On the other hand, there is descriptive and decision models that have existed for a long time. A descriptive model enables one to characterize the data, wherein, a country’s GDP and average life expectancy rate can be estimated. It is also exploratory in nature, where a customer provides the data and the problem is analyzed.  The customer is given an insight to the problem and then the decision model is used after which certain optimizations are proposed. The model has a target which is nothing but optimization.

# What is Logistic Regression?

[Logistic regression](https://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

# Modelling Techniques in Predictive Analytics

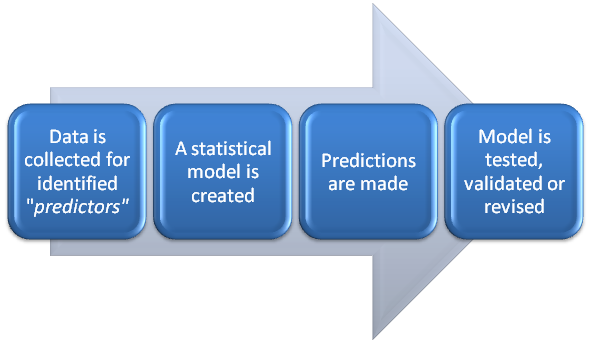


Models, in predictive analytics, represent meaningful relationships among variables, response or explanatory variables. *Response variable* refers to the quantity about which a query is made, while *explanatory variable* is the factor influencing the response variable. Explanatory variables are useful in predictive models – for observation, manipulation and control in relation to response. Patterns in historical and transactional data are studied for such variables to draw conclusions and make forecasts or identify opportunities.

**Predictive Modeling**

Predictive modeling is the creating, testing, validating and evaluating of a model to predict an outcome with a given input data.

A predictive model has *predictors*, i.e. variables likely to influence future outcomes. For instance, the gender, age group, location and purchase history of a customer adds up to the probability of a repeat transaction. These are ‘*predictors*‘.

[](http://ivyproschool.com/blog/wp-content/uploads/2015/06/Predictive-modelling.png)

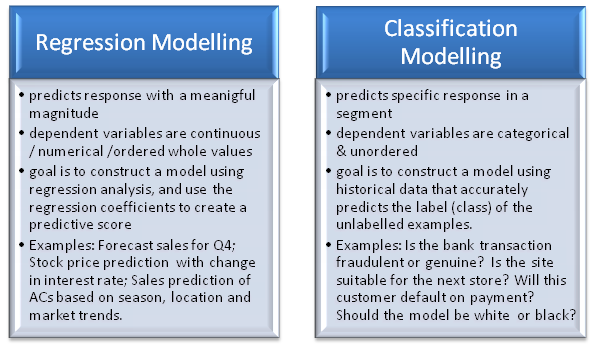
**What does a Predictive model do?**

* represents relationships in historical or/and real-time data
* looks for useful predictors or explanatory variables in the given data set
* fits models to subsets of data for evaluation of each predictor
* Makes observations, forecasts, classifications, predictions about future events.
* Assigns metrics for future actions

**Approaches of Predictive Modelling**

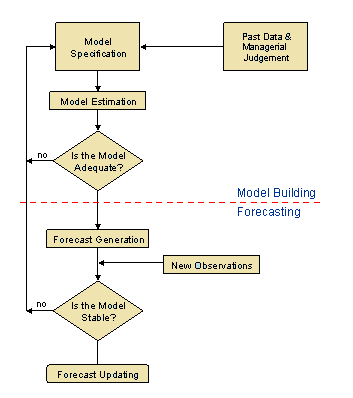
1. **Traditional –** driven by theories**;** uses classical or Bayesian methods of statistical inference; estimates parameters for linear predictors
2. **Data-adaptive** – driven by data; searches the data for subsets of useful predictors using machine learning methods; establishes non-linear relationships and interactions among variables
3. **Model dependent** – driven by specific model; uses the model/simulations to generate data, predictions or make recommendations; generated data is further compared with real data

**Common predictive models** are **regression** and **classification**.

[](http://ivyproschool.com/blog/wp-content/uploads/2015/06/Predictive-modelling-types.png)

**Common Statistical Models used in Predictive Analytics**

The techniques used differ for various applications. However, there are some fundamental statistical techniques, mathematical algorithms and neural network systems used in predictive modeling. Statistical models basically use mathematical equations to encode information which is extracted from the data and play a key role in data exploration.



### What are the types of predictive models?

Predictive model can be broadly classified into two categories: parametric and non-parametric. Parametric models make more assumptions and more specific assumptions about the characteristics of the population used in creating the model.

Some examples of parametric Machine Learning algorithms include:

* Logistic Regression
* Linear Discriminant Analysis
* Perceptron
* Naive Bayes
* Simple Neural Networks

Examples of popular nonparametric Machine Learning algorithms are:

* k-Nearest Neighbors
* Decision Trees like CART and C4.5
* Support Vector Machines

### **Statistical techniques and tools in Predictive Analytics**

* Linear regression
* Logistic regression
* Cluster analysis
* Analysis of variation (ANOVA)
* Chi-squared test
* Correlation
* Factor analysis
* Association rules
* Decision trees
* Time series
* Experimental design
* Bayesian theory – Naïve Bayes classifier
* Sampling
* Matrix operations
* K-nearest neighbor algorithm (k-NN)
* Pearson’s r

### **Commonly used Statistical models in Predictive Analytics**

**1. Logistic Regression:**

Logistic regression models the relation between a dependent and two or more independent variables (explanatory and response variables). It takes a look at how significant the relationship is between the variables. The probability (p) that event “1” occurs rather than event “2”. Where a good fit of the model is obtained, you can plug in the independent variable values for a new observation and predict if the dependent value will be 0 or 1.

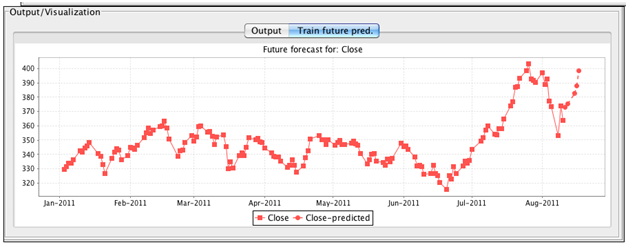
Examples in Predictive Analytics :

Banks – for building scorecards of customers applying for loans. The loan officer identifies characteristics that indicate probability of loan default, and further use this to build a scorecard of good and bad credit risks. Data of past, current and potential customers are used to execute a Logistic Regression Model. The model is leveraged to classify potential customers who have applied for loan, as good or bad credit risks. This uses binary logistic as the ‘dependent’ variable is dichotomous (loan default OR no default).

Education institutions – An engineering college would estimate enrolments of fresh students to determine cut-off marks and freeze admissions. A multiple logistic regression model is used to factor Class10, Class 12 and related AIJEE scores, distance from college, demographic information including stream preferences, historical data of student enrolments, to calculate probability of enrollment. The estimated model has to fit the data adequately to show the significance. Calculations can also be made to estimate the effect of how a single independent variable affects the likelihood of application.

**2. Time Series:**

The Time Series forecasting model is used to make predictions of future values based on previously observed / historical values. The two main goals are the identification of the phenomenon represented by the sequence of observations, and the forecasting of future values in the time series variable. The pattern of observed time series data is identified, described and integrated with other data. The identified pattern is further extrapolated to predict future events.

[](http://ivyproschool.com/blog/wp-content/uploads/2015/07/Model-eg.png)

Time Series predictive models are used to make forecasts where the temporal dimension is critical to the analysis. Typical application scenarios are demand prediction of a product during a particular month / period, estimation of inventory costs, forecast of train passengers for the next financial year, and so on.

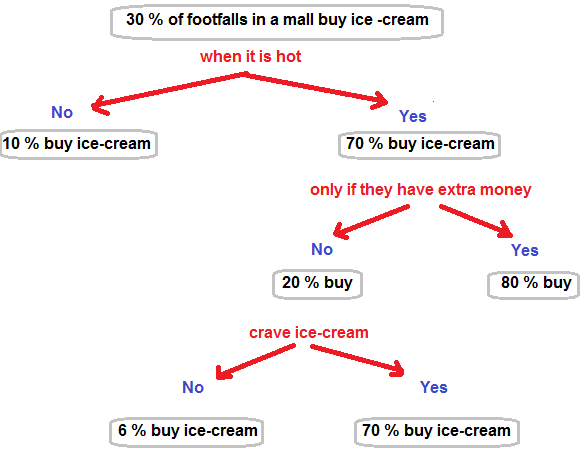
**3. Clustering:**

Clusters in the data are used for modelling predictions by grouping ‘like’ objects for a probability distribution. A model is hypothesized for each of the clusters to find the best fit of that model to each other. Clusters in customer behavior may be used for predictive modeling, i.e. behavioral clustering, to predict behavior or buying patterns of customers. Clusters in product segmentation may be used to predict what different categories of products customers are likely to buy. Algorithms auto-segment the objects based on several variables, to devise the cluster DNA. This is then leveraged for predictive insights.

Cluster models are used to predict demand of products (customer ordering baby clothes is likely to order diapers), brand preferences, predict efficacy of drug amongst a certain age group in clinical trials, predict stock market trends, identify groups of car insurance policy holders with a higher average claim cost, and more.

**4. Decision Trees:**

This statistical technique is a tree-like predictive model of decisions and possible consequences. Based on Boolean tests, specific facts are used to make general conclusions / decision points represented by nodes. Rules trace the series of paths from root to nodes, till an action is derived. Problems are structured as a tree with end nodes as branches, representing a specific event or scenario, or subject probability.

[](http://ivyproschool.com/blog/wp-content/uploads/2015/07/decision-tree.png)

A basic Decision Tree Modeling graph to predict how many buy ice cream because they crave for it, even if they don’t have extra money.

**5. Neural Network:**

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. [Neural networks](https://www.sciencedirect.com/science/article/pii/B9780080453828007292) consists of a network of nonlinear information processing elements called the neurons that are normally arranged in layers and executed in parallel. Neural networks are increasingly being used in areas of predictions and classifications, areas where statistical methods have traditionally been used. In fact, according to experts, the most commonly used artificial neural networks, called multilayer perceptions, are actually nonlinear regression and discriminant models. A neural network can approximate a wide range of predictive models with minimal demands on model structure and assumption.