PREDICTIVE MODELING

INDUSTRY APPLICATIONS

Website: www.analytixlabs.co.in

Email: info@analytixlabs.co.in

Disclaimer: This material is protected under copyright act AnalytixLabs©, 2011-2018. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions.

PREDICTIVE MODELS

Use models to predict what is likely to happen in the future, based on patterns in past data.
For example, models can predict the following situations:

- How likely it is that a customer will churn in the next quarter.
- Whether a customer will be a promoter of a service, or a detractor
- How valuable the customer is in terms of future revenue

Models can be used in the same way as business rules. However, while rules might be based on corporate policies, business logic, or other assumptions, models are built on actual observations of past results, and can discover patterns that might not otherwise be apparent. While business rules bring common business logic to applications, models lend insight and predictive power. The ability to combine models and rules is a powerful feature.

- **Training predictive models:** Predictive models must be trained to determine which data is useful and which data is not needed. When a model gives you accurate predictions, you are ready to use the predictive model for real time scoring.

- **Scoring a model:** To score a model means to apply it to some data in order to obtain a result or prediction that can be used as input to decisions.

- **Deployment:** You can deploy the application to a testing environment or to a real-time production environment, such as a call center or a website. You can also deploy it to contribute to batch processing.

DATA SOURCES

You need the following types of data in the modeling process:

Historical or analytical data:
To build the model, you need information about what to predict. For example, if you want to predict churn, you need information about customers such as their complaints history, number of months since they upgraded their plan, sentiment score, demographic history, and estimated income. This is often referred to as historical data or analytical data, and it must contain some or all of the fields in the project data model, plus an additional field that records the outcome or result that you want to predict. This extra field is used as the target for modeling.

Operational or scoring data:
To use the model to predict future results, you need data about the group or population that you are interested in, such as incoming claims. This is often referred to as operational data or scoring data. The project data model is typically based on this data.

TRAINING PREDICTIVE MODELS

Predictive models must be trained to determine which data is useful and which data is not needed. When a model gives you accurate predictions, you are ready to use the predictive model for real time scoring.
You use a training data set to build the predictive model and a test set of data to validate the model that was created with the training set.

SCORING A MODEL

To score a model means to apply it to some data in order to obtain a result or prediction that can be used as input to decisions.

Depending on the application, the scoring results can be written to a database table or flat file, or used as inputs to the segment, selection, and allocation rules that drive decisions in an application.

CREATING BUSINESS RULES

The insights gained through predictive modeling can be translated to specific actions.

You can combine predictive models with rules to allocate offers in accordance with business goals. This is done using a combination of selection and allocation rules that are based on the output from predictive models.

The steps you take are:
Define possible actions: If a customer is not happy with a service, what should you do about it.
Allocate offers: Which types of customers are the best candidates for which offers.
Prioritize offers: Prioritization determines which offers a customer will receive.

DEPLOYMENT

You can deploy the application to a testing environment or to a real-time production environment, such as a call center or a website. You can also deploy it to contribute to batch processing. For example, a model can be automatically updated at regularly scheduled intervals as new data becomes available.

PREDICTIVE MODELS IN THE TELECOMMUNICATION INDUSTRY

A number of predictive models are provided in the Telecommunications industry.

The following models form the basis of the predictive models in the Telecommunications industry:

Churn model: Customers likely to churn from the current list of active customers can be predicted.
Customer Satisfaction model: Customer satisfaction is determined by the net promoter score.
Association model: Customers can be profiled and assigned to segments.
Response propensity model: You can determine the correct channel to reach the customer, and the probability that the customer will respond.

- **Predicting churn**: Churn is the measurement of subscribers who ended their contract or services. The objective of the Churn Prediction model is to predict the customers likely to churn from the current list of active customers.
  The inputs for the example Churn Prediction model are complaint history, number of months since the customer upgraded the plan, sentiment score, customer demographic history, and estimated income.

  Data preparation for churn prediction starts with aggregating all available information about the customer. The data that is obtained for predicting the churn is classified in the following categories:
    - Transaction and billing data, such as the kind of services subscribed, and average monthly bills.
    - Demographic data, such as gender, education, and marital status.
    - Behavior data, such as complaints data and price plan migration data.
    - Usage data, such as the number of calls and the number of text messages sent.

  Data is filtered for modeling in two stages:
    1. Data not relevant to some customers.
    2. Variables that do not have adequate predictive significance.

  A CHAID algorithm is used to predict churn. A CHAID algorithm generates decision trees. A decision tree model is selected over logistic regression because the rules that come out of the decision tree help to understand the root cause of churn better.

  The sentiment score is derived from the customer comments text and is an important predictor of churn. Sentiment score considers both the current sentiment score and historical sentiment score.
  Other important predictors that are identified during the data understanding and modeling phase are estimated income, number of open complaints, number of closed complaints, time since the last plan upgrade, and the education level of the customer.

  Along with the probability of churn occurring, the propensity to churn is calculated by the model.

- **Customer satisfaction:** Customer satisfaction in the Telecommunications sample is determined by the Net Promoter Score (NPS).

  The Net Promoter Score is based on the perspective that every company's customers can be divided into three categories:

- Promoters are loyal enthusiasts who keep buying from a company and urge their friends to do the same.
- Passives are satisfied but unenthusiastic customers who can be easily wooed by the competition.
- Detractors are unhappy customers who are trapped in a bad relationship with the company.

The Net Promoter Score is obtained by asking a set of customers a single question: "How likely is it that you would recommend our company to a friend or colleague?" Customers are asked to answer on a 0 - 10 rating scale. Based on the score that they provide, they are categorized as Promoter (if the score is 9 or 10), Passive (if the score is 7 or 8), or Detractor (if the score is 6 or less).

The objective of the Net Promoter Score model is to identify the distinguishing characteristics of the customers who fall into the three categories. The net promoter score model is then used to predict which category a customer would fall into, without asking the question "How likely is it that you would recommend our company to a friend or colleague?" This model helps to dynamically track the change in the Net Promoter Score of a customer.

Historical data comes from a sample of customers who answered the question. Customers for whom there is no score are considered to be operational data, whose satisfaction group needs to be predicted for the first time. The Customer Satisfaction model can be used to predict scores for customers who do not have a net promoter score.

The sentiment score, along with the number of open complaints, employment status, and estimated income, are identified to be the key variables that affect the prediction of satisfaction group. The sentiment score is focused on capturing the negative sentiments across various attributes, such as network and service. A sentiment score of zero means that the customer has not expressed any negative sentiment. A sentiment score of two means that the customer has expressed negative sentiment in two predefined categories. Six categories were defined, and so the maximum sentiment score is 6.
The sentiment score that is used in the example database is an average value of the most recent sentiment score calculated and the previous sentiment score of the same customer. Where a customer expressed negative sentiment on a single category, and then expressed multiple positive comments, the sentiment score would be mildly negative, although close to zero. For the purposes of satisfaction modeling, to avoid categorizing the customer as mildly negative, sentiment scores less than 0.6 are rounded to zero.

- **Assigning offers:** An Association model is used to assign the right offer to a customer. It uses the customer's segment (for example, Platinum) and predicted net promoter score group (for example, Promoter) to determine an offer (for example Phone Plan).

  Segmentation is the process of profiling customers into groups with similar demand characteristics.

- **Targeting offers to customers with the response propensity model:** It is important to target offers to the correct customers, through the correct channel.

  The Response Propensity model determines the correct channel to reach the customer, and determines the probability that the customer will respond.

  You can use the results of this model to target customers who are likely to respond because they are above a certain threshold, or ignore customers who are likely to result in a minimum profit.

The input for the model is customer demographic information, billing history, customer lifetime value, churn score, net promoter score, and tenure.

The customer's previous offer response data can be used as the input for the current model. The historical data on which interaction points the customer has responded to an offer is taken and the model is trained based on that data.

- **Telecommunications models in Analytical Decision Management**: You can combine predictive models with rules to allocate offers in accordance with business goals. You do this by combining selection and allocation rules that are based on the output from predictive models.
  There are two main steps:
  - Define and allocate offers to determine which offers a customer is eligible for.
  - Prioritize offers to determine which offers a customer receives.

3. PREDICTIVE MODELS IN THE RETAIL INDUSTRY

The following models form the basis of the predictive models in the Retail sample.

Customer Segmentation model: Customer segmentation involves profiling customers through demographic segmentation, online behavior segmentation, and buying behavior segmentation.

Market Basket Analysis model: Market basket analysis allows retailers to gain insight into the product sales patterns by analyzing historical sales records and customers' online browsing behavior.

Customer Affinity model: You can determine customer affinity towards product lines by understanding the customer demographic information, purchase information, and browsing information.

Response Log Analysis model: Response log analysis captures the response of customers compared with the recommendations that come from IBM® Analytical Decision Management business rules.

Price Sensitivity model: Price sensitivity is the extent to which the price of a product affects a customer's purchase decision. The degree of price sensitivity varies from customer to customer, and from product to product.

Inventory-Based Suggestion model: The Inventory-Based Suggestion model identifies the products that have excess inventory and then makes real-time recommendations to the customers based on the combination of category affinity and excess inventory.

- **Prepare data for a retail solution:** Before data can be used in predictive models, process the online browsing data

Pre-processing online data:
Use a data pre-processing stream to process the browsing behavior data from an online system and load it into a database table in a format that is suitable for modeling analysis. Processing can be done in batch processing mode.

To obtain the online browsing behavior data, deploy a web analytics solution. Web analytics tools allow the flexibility to export the browsing behavior data in different formats such as a comma-separated file, which can then be consumed for further analysis. Data might include information such as products browsed, products put into a cart, products abandoned, products purchased, pages viewed by customers, and the product category.

Category affinity target determination:
Affinity analysis is a data analysis and data mining technique that discovers co-occurrence relationships among activities that are performed by the customer to understand the purchase behavior of customers. The customer can have both an online and a physical presence. During online shopping, customers can browse, search, and view pages for different products before they make a purchase. The purpose of the Category Affinity model is to get information about product lines in which a customer is interested by understanding the online and buying behavior.
Use a category affinity target determination stream to process customer online and in-store historical transactional data to find out the following information:

- Products browsed
- Products purchased
- Products abandoned
- Products that are put in a cart
- Onsite searches
- Page views

The stream should process each activity separately so that you can prioritize the order of activity as purchase, browse, search, and page views.

Data should be processed in two steps:

1. Aggregation of the data should be done at the product line level so that you can get the number of items per product line for each customer activity.
2. Identify whether a customer has an affinity for a particular product line by comparing:
   - The number of items that are purchased in a category, with
   - The average number of items that are purchased by the total population.

Some product lines are more likely to be purchased in large quantities while some other products would be purchased in small quantities. For example, a customer might buy many writeable DVDs but they might buy a computer once in three years. If you take the number of items or the value of items, the Category Affinity model would show a bias to a select few products.

Customers are likely to search products with different names, and might use related keywords or visit related pages. The pre-processing stream can process available data to get the corresponding product line information and derive the volume of items for each product line under this category.

When a customer searches a product with an exact product line name, a weight of one is assigned to that product line. However, when a customer searches for a super category, the number of product lines in that super category is determined and all the product lines in that super category are assigned equal weight. For example, a customer searches by the category named Consumer Electronics. It is not possible to know

Which product line the customer searched for because Consumer Electronics contains three product lines: Computers, MP3, and Smartphone. In such cases, all three product lines have a weight of 1/3 = 0.333333333.

You can get more insight into customers from their purchase behavior than from their browsing behavior. Consider browsing behavior only when there is no purchase information. Likewise, if there is no browsing behavior, consider searching behavior, and if there is no searching behavior, consider page view information.

- **Defining customer segments:**
  When you define customer segments, you profile the customers into groups that have similar demand characteristics.

  You profile customers based on demographics, online behavior, and buying behavior to help to provide the right offer for the customer at the right time.

Demographic segmentation:
Demographic segmentation is based on age, gender, marital status, number of members in household, education, profession, and income. To get meaningful segments out of the data, missing information is derived based on the other variables. Continuous variables such as age and income are tiled into smaller number of groups.

When there are multiple variables to consider, clustering can be challenging. K-Means clustering can be used to cluster the customer base into distinct groups. Instead of trying to predict an outcome, K-Means clustering tries to uncover patterns from the demographic data that is provided as input. The clusters are formed in such a way that each customer within the segments is similar and different from customers present in other segments. Multiple iterations of K-Means clustering on the customer base are performed to arrive at six clusters to be used for targeting campaigns. Education, income, and marital status are the top three variables that determine the cluster in which a customer is categorized.

Online behavior segmentation:
Before online behavior segmentation can begin, the data must be prepared. Data preparation is described in Prepare data for a retail solution. This data is then aggregated to reveal trends in a particular session, and across sessions for an individual customer, and the total population.

The online browsing history is taken, and based on the aggregate information that is prepared in the data preparation stage, a cluster model is developed. The clusters that are formed by clustering method depict the qualities of customers in the various stages of the online purchase funnel.

Buying behavior segmentation:
The purchase history of the customer, both online and in-store, is collected. Extra details that relate to the products purchased are also collected. The extra details might include: was the item purchased when there was an offer? What was the discount? What is the margin from selling the product? This information is used to derive multiple variables such as the average purchase value and average number of items that are purchased, across both channels and during discount and regular shopping.

The past purchase behavior helps to predict how likely a customer is to react to various kinds of offers. The past purchase behavior also helps to predict the kind of campaign that would be more suitable for the customer. Based on the online and in-store purchases that are made by the customers, and whether the purchases were made during an offer or at another time, a K-means clustering model is used to derive the various segments. The model identifies two clusters for primarily online customers, and three clusters for primarily in-store customers. For both online and in-store clusters, one segment is identified as offer hunters. Online offer hunters typically purchase high value items and in-store offer hunters typically purchase low value items. Non-offer hunters typically purchase high value items.

- **Market Basket Analysis:**
  Market Basket Analysis allows retailers to gain insight into the product sales patterns by analyzing historical sales records and customers' online browsing behavior.

  Market Basket Analysis is used to increase marketing effectiveness and to improve cross-sell and up-sell opportunities by making the right offer to the right customer. For a retailer, good promotions translate into increased revenue and profits. The objectives of the market basket analysis models are to identify the next product that the customer might be interested to purchase or to browse.

Market Basket Analysis works on relating products in the historical transactions of a retailer. Association rules are generated by using the frequency of a particular item along with the combination of items. Rules with higher lift, confidence, and support are selected for deployment.

Market Basket Analysis for an online retailer requires two types of data: sales transaction data, and the customer's online browsing behavior data.

Online data that was prepared in the data processing step, provides the information on online purchase and browsing. For more information, see Prepare data for a retail solution. The in-store transaction tables provide the information on physical store purchase.

An Apriori algorithm is used to find association rules between different product categories. Market Basket Analysis is done separately to find associations between products that are browsed, products purchased online, and products purchased in-store. Browsing behavior data is aggregated to get all the products purchased by a customer and all the products browsed by a customer. The Apriori algorithm is then applied on the aggregated data to find the association between different product categories for the products that are purchased and the products that are browsed.

- **Determine Customer Affinity:**
  You can determine customer affinity towards product lines by understanding the customer demographic information, purchase information, and browsing information.

  The input for the Category Affinity Model is the output of the Category Affinity Target Determination model, the Customer Segmentation model, and online transactional data. The outputs from the models are stored in the intermediate database tables, and contain information about the product lines that customers are most interested in, the customer segments, and the customer market basket. This input is used by a logistic regression algorithm. Logistic regression is a statistical technique for classifying records that are based on values of input fields. A multinomial model is used because the target has multiple product lines.

- **Response log analysis**
  A response log captures the response of customers compared with the recommendations that comes from the above model

  A response log gives a view of how many recommendations are taken up by customers in the form of offers that are accepted or offers that are declined. The objective of the Response Log Analysis model is to discover the following information:
  - The patterns of customers who are converted to Buyers.
  - Which rules triggered high conversion rates to determine high impact business rules.

  The following models are used:
  - Self-Learning Response Model (SLRM) algorithm
  - Bayesian Network algorithm

  Response log data is captured by the Response service. The Response service logs all the customer responses to IBM SPSS Collaboration and Deployment Services system tables in the form of XML tags. The log contains customer responses such as offers accepted, offers made, customer demographics, actual profit, IBM Analytical Decision Management rules for the offers that are accepted, and other metadata.

A similar method can be used to log product feedback that is given by customers in text format against each product line. This data is queried by XQuery, a query and functional programming language that queries collections of XML data. The data is then loaded into a view and used as a source of data for modeling.

The Self-Learning Response Model (SLRM) algorithm is used to predict the best offers to customers by using past responses to recommendations, and customer demographics. By using the SLRM node, you can build a model that continually updates, or re-estimates, as a data set grows without having to rebuild the model using the complete data set. This model predicts which offers are most appropriate for customers and predicts the probability of offers being accepted. The model predicts the best three offers for a customer. The model also analyses IBM Analytical Decision Management rules to determine which are the most effective rules?

- **Inventory-based suggestion**

Retailers often have the problem of excess inventory, where the value of the inventory depreciates rapidly due to the products becoming outdated. To prevent this problem, retailers use offers in order to clear the excess inventory. The Inventory-Based Suggestion model identifies the products that have excess inventory and then makes real-time recommendations to the customers based on the combination of category affinity and excess inventory.

The input to the model is transaction data from the online and physical store, and product details, including the current inventory. The output is the product name, and price and cost data to be made available to the customer.

The modeling techniques that are used are time series modeling.

Forecasting inventory
You can predict the demand for products a week in advance.

Purchases that are made by the customers both in-store and online are aggregated by day, giving the information on demand for products daily. This information is used as an input for time series modeling. In the Time Series model, you can build the model for each of the products depending on their individual characteristics.

Inventory cost analysis
You can calculate excess inventory and the holding cost. Excess inventory is determined by using the current stock, forecasted demand for the period in consideration, and the variation in demand.

> Excess Inventory = Current Stock – Forecasted Demand – score required for the service level * Expected Variance in demand

Excess inventory is calculated a week in advance. The current stock is taken from the PACK SIZE variable in the product table. The forecasted demand for the next seven days is given out by the time series model. The standard deviation for a single day is calculated by using the aggregation, which gives the values for standard deviation for all products. To get the variance over the length (seven days) the variance is multiplied by SQRT(7). The variance over a longer duration is an addition of the variance expected for each day, and standard deviation is the square root of variance. The holding cost is then taken to be 25% of the cost of all the products that are in excess inventory.

Real-time recommendations
When a customer's data is available, the product line to which the customer has the highest affinity is selected in real time. If there is a product in that product line that has excess inventory, that product is recommended to the customer. If the customer's affinity is with a product line that does not have excess inventory, then the default product, which has the highest holding cost across all product lines, is recommended to the customer.

- ## Deployment of models in the Retail sample

Scoring the data based on the below predictive models that are described in the Retail case study.

- Online behavior segmentation.
- Browsing Market Basket Analysis at a product level by customer.
- Buying Market Basket Analysis at a product level by customer.

To perform Market Basket Analysis, you must look up the price information in the product table or the product that comes out of the Market Basket Analysis recommendation

- ## Use of Retail case study models

Retail Promotions can be designed for campaign management in a retail scenario. It is shared by two Analytical Decision models, one model for online promotions, and one model for in-store promotions.

Online promotions
For online promotions, campaigns are designed to target specific customer segments.

The data that is used by the application includes demographic, behavioral data, and purchase history, and attributes such as segment membership and category affinity that are derived from predictive models.

In-store promotions

- For in-store promotions, campaigns are organized around business objectives, such as reducing inventory through special promotions or rewarding the most loyal customers.

Input to the Analytical Decision Management
We can use output of the following predictive models as input:

- **Category Affinity model output**
  The probability that the customer likes a particular product line.

- **Segmentation model output**
  The demographic, online behavior, and buying behavior segment outputs.

- **Market Basket Analysis output**
  The output of market basket analysis that is based on browsing and purchase.

4. PREDECTIVE MODELS IN THE INSURANCE INDUSTRY

The following models form the basis of the predictive models in the Insurance industry:

Segmentation model

Customers are segmented based on their financial sophistication. This model enables the insurer to sell insurance policies that are appropriate to the customer.

Churn Prediction model

Predicts a customer's propensity to churn by using information about the customer such as household and financial data, transactional data, and behavioral data.

Customer Lifetime Value model (CLTV)

Predicts customer lifetime value. CLTV is based on the revenue that the customer brings in to the company, the cost of maintaining the policies, the cost of retention, and the likelihood of the customer surrendering the policies soon.

Campaign Response model

Predicts the probability that customers will respond to targeted offers so that you send out offers only to the customers whose propensity to respond is above a particular threshold.

Auto Churn model

Predicts the customers likely to churn from the current list of active customers. This model considers the Auto policy customers only.

Life Stage Segment model

Groups customers based on their current life stage, which would help in recommending the right insurance policy to the customer based on their current life stage segment.

Buying Propensity model

Identifies the life policies that are mostly bought by customers belonging to each life stage segment as defined in the life stage segment model.

Data Processing Stream

Transforms and aggregates data obtained from the customer's visit to the insurer's website so that it can be used to define rules for recommending the right insurance policy to each customer.

Insurance Policy Recommendation model

Recommends the correct insurance policy by considering information such as the customer’s web activity data and also the life insurance policy buying propensity for the given customer’s life stage segment, based on historical data.

Social Media Analytics (SMA) model

Extracts customer life stage event information from customer’s social media posts.

Sentiment scoring model:

Extracts the sentiment score from customer comments that are captured while recording customer complaints.

- Typical **Data used in the Insurance Analytics**
  The Insurance Company runs a multi-line business. The following types of data are used.

  Customer master data

  This includes customer's demographic data, employment and income data, and information about the household as well. POLICYHOLDER and HOUSEHOLD tables capture most of this data. Typically, Master Data Management systems are the source of customer master data.

  Customer policy data

  This includes aggregated customer information, such as the number and types of policies owned by the customer, total premium being paid by the customer, average claim amount, tenure of the customer, number of complaints, number of claims, and customer sentiment data. POLICYHOLDER_FACT and POLICY_FACT tables capture most of this data.

  Customer transaction data

  This includes data about all the customer transactions such as the policies purchased, their inception and maturity/renewal date, data related to all of the complaints made by the customer in the past, and also data related to all of the claims made by the customer. POLICIES, CLAIMS, COMPLAINTS, COMPLAINT_DETAILS tables contain this data.

  Customer Social Media Data

  Apart from the customer data that is available within the enterprise, insurance organizations may also want to get insights from external sources of data. For example, the social media channels where customers post comments about their experiences with their insurers, as well as about their needs and life-events that can potentially lead to an opportunity to sell appropriate insurance products. SMA_DATA and SMA_DATA_ANALYSIS tables captures such external data, as well as the summarized analysis of this social media data.

  Customer web browsing data

  Many insurance organizations today allow their customers to buy or explore their insurance products online through their websites. Technology makes it possible to track customers' activities on their websites, giving them vital insights about the customers' current interest in specific insurance products. Web Analytics tools can be used to analyze customers' website activities and use this information along with other customer data to make the right recommendations to the customers at the right time. ACTIVITY_FEED_DATA, ONLINE_BROWSING_HISTORY and ONLINE_BROWSING_SUMMARY tables contain customer's web activity data.

- **Define customer segments in the Insurance Industry**

  When you define customer segments, you profile the customers into groups that have similar demand characteristics. In the Insurance Industry, customers are profiled based on their financial sophistication.

  Customers are segmented into financially sophisticated, and novice categories. This means that insurers can target each segment with cross-selling insurance policies that are appropriate to increase the effectiveness of cross-sell campaigns.

The inputs for the Segmentation model are customer master data and customer policy data, specifically:

- Demographic data: age, gender, marital status, and employment status.
- Insurance policy related data: insurance lines, policies, premiums, tenure, and insurance score.
- Financial data: income, retirement plans, home ownership status, vehicle ownership.

These inputs are aggregated. As each record is read, based on a distance criterion, the cluster algorithm helps to create clusters

- **Predict churn in the Insurance Industry**

The Churn prediction model predicts a customer's propensity to churn by using information about the customer such as household and financial data, transactional data, and behavioral data.

The inputs for the Churn prediction model are customer demographic data, insurance policies, premiums, tenure, claims, complaints, and the sentiment score from past surveys.

Data preparation for churn prediction starts with aggregating all available information about the customer.

The data that is obtained for predicting the churn is classified in the following categories:

- Demographic data, such as age, gender, education, marital status, employment status, income, home ownership status, and retirement plan.
- Policy-related data, such as insurance lines, number of policies in the household, household tenure, premium, disposable income, and insured cars.
- Claims, such as claim settlement duration, number of claims that are filed and denied.
- Complaints, such as number of open and closed complaints.
- Survey sentiment data. Sentiment scores from past surveys are captured in the latest, and average note attitude score fields. The note attitude score is derived from customer negative feedback only. If the note attitude is zero, the customer is more satisfied while as the number increases, satisfaction level decreases.

Any classification algorithm can be used to predict churn. The output of the model also provides the most important predictors influence customer to churn. For example, the most important predictors could be HOUSEHOLD_TENURE, LATEST_NOTE_ATTITUDE, and NUMBER_OF_POLICIES_IN_HOUSEHOLD.

- **Understand Customer Lifetime Value (CLTV)**

The Insurance sample uses Customer Lifetime Value (CLTV) to understand customer profitability.

CLTV is a commonly used approach to determine how much each customer is worth in monetary terms. It helps insurers to determine how much money must be spent to acquire or retain a customer. CLTV is defined as the expected net profit a customer will contribute to the business over time. Advanced analytics provide new insights on how customer lifetime value can be calculated.

The inputs for this model are customer demographic data, policies, premiums, tenure, policy maintenance cost, complaints, and customer survey sentiment.

CLTV is determined by the margin amount that the customer contributes each month and the probability that the customer might churn in any month. Customers with higher margin amount and a lower probability to churn have a high CLTV.

CLTV is derived by the following formula:

$$\sum_{i=0}^{N} \frac{NetProfit * C_i}{(1+d)^t}$$

Ci= probability for customer *i* to generate revenue in time $_t$
N = total number of periods
d = monthly discount rate
t = time of cash flow

The probability *Ci* can be estimated by using Cox regression. Cox regression is a method for investigating the effect of several variables upon the time a specified event takes to happen. In the context of an outcome such as churn, this is known as Cox regression for survival analysis.

CLTV is calculated by considering the following:

- When the Cox model is scored, the customer's past 'survival' time is considered, and the churn probability is predicted for one to five years.
- The NetProfit value is derived by using following expression.

```
NET_PROFIT =(TOTAL_PREMIUM - MAINTENANCE_COST)*12
```

- The customer lifetime value is derived as follows:

```
CLTV =(NET_PROFIT * $C_1$/  (1 + 0.12))
                        + (NET_PROFIT * $C_2$/  (1 + 0.11) ** 2)
                        +  (NET_PROFIT * $C_3$ /  (1 + 0.1) ** 3)
                        +  (NET_PROFIT * $C_4$ /  (1 + 0.09) ** 4)
                        +  (NET_PROFIT * $C_5$ /  (1 + 0.08) ** 5)
                        + (NET_PROFIT * POLICYHOLDER_TENURE)
```

Where $C_{i=(1,2,3,4,5)}$ is the renewal probability, which is the future value that a customer can bring. The last item is the historical/current value of one customer.

- The CLTV values are further classified into Low, Medium, and High categories by using the following calculations:

```
CLTV_CAT = if CLTV <=30083.625  then 'LOW'
               elseif CLTV > 30083.625 and CLTV <= 46488.000000000007 then 'MEDIUM'
               elseif CLTV > 46488.000000000007 then 'HIGH'
               else 'LOW'
               endif
```

- **Predict customer response to a campaign**

  Targeting offers to the correct customers is an important part of promotion planning and campaign design. The Insurance industry uses the Campaign Response model to predict the probability that the customer will respond to targeted offers.

  The Campaign Response model helps in sending out offers only to the customers whose propensity to respond is above a particular threshold.

  The customer's previous offer response data is the input for the model, and the model is trained based on that data.

  The Decision List algorithm is used to identify the characteristics of customers who are most likely to respond favorably based on previous campaigns. The model generates rules that indicate a higher or lower likelihood of a binary (1 or 0) outcome. The Campaign Response model considers only those customers who are currently with the insurance company, and not those customers who churned.

- **Predict churn for auto policy holders**

  The Insurance sample uses the Auto Churn prediction model to predict the customers likely to churn from the current list of active customers that hold auto policies.

  The classification algorithm can be used to predict churn. This model is similar to the Churn model, except that this model only considers auto policy data for predicting churn.
  You can also identify most important predictors of churn. Example: the last survey sentiment score (LATEST_NOTE_ATTITUDE), and CLAIM_SETTLEMENT_DURATION.

- **Group customers based on their current life stage**

  The Insurance sample uses the Lifestage Segment model to group customers based on their current life stage.

  The model uses simple rules to get the current life stage segment of customer. Some examples of defined segments are:

  - Newly married.
  - Young family.
  - Young and affluent.
  - Single.
  - Divorced.

- **Identify the life policies bought by life stage segments**
  The Insurance sample categorizes customer buying propensity for life policies into life stage segments using the Apriori model.

  The Apriori model is an association algorithm that extracts association rules from data. The algorithm uses the past insurance policy purchase data and provides the buying propensity score for each policy at customer life stage segment level. The output is processed further into a summary format, which is then used to deliver appropriate offers to customers.

- **Recommend the correct insurance policy**

The Insurance industry uses the Insurance Policy Recommendation model to recommend the correct insurance policy for customers.

The Insurance Policy Recommendation model compares insurance policies that are browsed by the customer with the list of insurance policies that have a higher buying propensity for the customer's life stage segment. The model recommends the correct insurance policy according to this data.

Transform and aggregate customer data: Transform and aggregate data about customer activity on the insurer's website. To obtain the online browsing behavior data, deploy a web analytics solution.
Extract life stage event information from social media posts: You can use social media to get valuable insights about customers. We can use Text Analytics to read the social media data, and to extract life stage event information. Some examples of life stage events are new baby born, new job, new house, birthday, marriage, and so on. This information is used to help recommend appropriate insurance policies to customers.
Extract sentiment scores from customer complaints: Customer interactions with call center agents can be a source of valuable data to determine customer satisfaction levels. We can use the Sentiment Scoring model to extract the sentiment score from customer comments that are captured while recording customer complaints.

The input is customer complaint details. The Text Analytics model reads the customer complaint details, and extracts meaningful words and concepts from the information. Negative concepts are used to derive the sentiment score. The sentiment score reflects the count of negative words used by customers while making the complaints. Example concepts are "bad", "not accessible", "slow", "wrong".

- **Insurance data model:** The historic data that is used for predictive modeling in Insurance Industry. The following table describes some of the data columns that are part of the Database Views.

NAME	**DESCRIPTION**
AGE	The age of the policy holder.
CLTV	Customer Lifetime Value.
EDUCATION	The education level of the policy holder.
EMPLOYMENT_STATUS	The status of the person’s employment.
GENDER	The person’s sex or gender.
INCOME	The policy holder’s annual income.
MARITAL_STATUS	The marital status of the person.
MAINTENANCE_COST	The cost of maintaining this policy.
MONTHS_SINCE_POLICY_INCEPTION	The number of months since the policy holder started the policy.
MONTHS_SINCE_LAST_CLAIM	The number of months since the policy holder filed the last claim.
NUMBER_OF_CLAIMS_DENIED	The number of claims that were denied.
NUMBER_OF_CLAIMS_FILED	The number of claims that are filed.
CLAIM_SETTLEMENT_DURATION	The time, in days between the date when the claim opened and the date when the claim was closed, and the customer satisfaction confirmed, based on the status of the claim.
NUMBER_OF_COMPLAINTS	The number of complaints the policy holder has submitted.

NAME	DESCRIPTION
NO_OF_CLOSED_COMPLAINTS	The number of complaints that have been closed.
NUMBER_OF_OPEN_COMPLAINTS	The number of complaints that are open.
LATEST_NOTE_ATTITUDE	The noted attitude of the last communication.
AVG_NOTE_ATTITUDE	Average communication note attitude.
NUMBER_OF_POLICIES	The number of policies that the policy holder has.
POLICYHOLDER_ID	Any value without business meaning that uniquely distinguishes each occurrence of this entity.
POLICY_ID	Any value without business meaning that uniquely distinguishes each occurrence of this entity.
POLICY_TYPE	Indicates the policy type, for example, fixed term and flexible term.
VEHICLE_OWNERSHIP	Indicates whether the policy holder owns a vehicle or not.
VEHICLE_TYPE	The vehicle type.
VEHICLE_SIZE	The vehicle size.
HOME_OWNERSHIP_STATUS	The tenancy status of residence.
INSURANCE_LINES	The number of types of insurance products held by the policy holder.
INSURANCE_SCORE	The insurance score, based on the credit score, as well as other factors such as claim filing history.
LIFE_CUSTOMER	Indicates whether a customer owns a life insurance policy.
NON_LIFE_CUSTOMER	Indicates whether a person owns a non-life insurance policy.
NUMBER_OF_CHILDREN	The number of children of the policy holder.
NUMBER_OF_INSURED_CARS	The number of insured cars insured with the insurance company.
POLICYHOLDER_TENURE	The number of years the policy holder has been a customer with the insurance company.
TOTAL_PREMIUM	The total premium on all the policies paid by the policyholder to the insurance company.
RETIREMENT_PLAN	The name of the retirement plan owned by the policyholder
HOUSEHOLD_NUMBER_OF_CHILDREN	The number of children in household
HOUSEHOLD_NUMBER_OF_INSURED_CARS	The number of insured cars in the household that are insured with the insurance company.
NUMBER_OF_POLICIES_IN_HOUSEHOLD	The number of policies held by the household.
HOUSEHOLD_TENURE	The number of years that a household has been classified within its customer status.
HOUSEHOLD_PREMIUM	The total premium paid by the household to the insurance company.
HOUSEHOLD_DISPOSABLE_INCOME	The amount of money that the household can spend after having paid all the fixed expenses such as rent, mortgage repayment and so on.

5. PREDICTIVE MODELS IN THE ENERGY AND UTILITIES INDUSTRY

A number of predictive models are provided in the Energy and Utilities Industry.

Credit Rating Model: Predict the probability that customers will miss payments.
Demand Response Program Acceptance: Predict the probability that customers will need demand response programs.
Recommended Rate Plan: Recommend the most profitable Rate Plan.
Sentiment: Determine the sentiment score for customers based on social media information, such as Twitter.
Satisfaction: Measure customer satisfaction as determined by the net promoter score.

- **Predict credit rating**
  The Credit Rating model predicts the probability that customers will miss payments.

  The Credit Rating model can be developed to classify customers based on credit history, annual income, ownership, family members, and so on. Any classification technique like logistic regression algorithm is applied to read past payment history and customer data to evaluate customers' ability to pay their bills.

- **Predict the need for demand response program assistance**
  The Demand Response Program Acceptance model predicts the probability that customers will need demand response programs.

  A Classification algorithm reads the past history of demand response programs and customer demographic information and uses this information to predict whether customers will need demand response programs assistance in the future.

  A demand response program encourages customers to reduce their energy use at times of peak demand to save money or gain incentives.

  Recommend the right rate plan
  The Recommended Rate Plan model in the Energy and Utilities sample recommends the most profitable rate plan.

  Average Energy and Utilities usage is considered by customer, and the bill amount is calculated for all available rate plans. The current rate plan bill totals are compared with all other rate plan bill totals, and from this comparison, the most profitable rate plan is selected.

- **Understand customer sentiment**
  The Sentiment model in the Energy and Utilities sample determines customer sentiment.
  Text Analytics model mines unstructured data to establish the customer sentiment. The data comes from social media sources such as Twitter. The data is categorized into areas such as outages, power line issues, customer service, and bill issues. A sentiment score is determined, based on the set of categories that is defined in the Text Analytics node.

6. PREDICTIVE MODELS IN THE BANKING INDUSTRY

Q: Briefly, how can analytics be used in the banking industry?

Ans: The banking, financial services, and insurance (BFSI) industry was an early adaptor of analytics. Today, almost all bank and financial institutions use analytics aggressively, and the industry is full of some very meaningful use cases that have helped it hugely.

Analytics tools give banks insights into the personal habits of its customers, allowing them to promote offers accordingly. Analytics is also used by banks to reduce the chances of money laundering, by identifying suspicious activity, such as moving money to multiple accounts, finding large single-day cash deposits, the opening of a number of accounts in a short period of time, or sudden activity in long-dormant accounts.
Using analytics, a bank is also able to keep track of the credit histories of customers and can distribute loans accordingly.

When salespeople are pitching a loan to a client, a bank tries to find out the background of the customer and what the likelihood of his/her taking a particular loan is. Banks also use analytics to increase customer loyalty and reduce loan prepayments due to refinancing with other institutions.
To follow the AML (anti money-laundering) guidelines around fraud analysis, banks deploy tools that can identify complex schemes/transactions.

The tool provides link analysis to investigate financial similarities between apparently unrelated accounts, to detect money being moved across accounts. It helps in establishing connecting patterns in potentially fraudulent transactions, by scanning a history of transactional data.
Banks are also applying their data models to education loans, automotive loans, housing loans, and loans to small- and medium-sized enterprises (SMEs), to try and reduce the percentage of those loans going bad.
For instance, in the case of an education loan, banks combine data from their bad loans, income tax departments, and credit ratings agencies to identify suitable candidates and then send them reminder messages on Facebook.
Banks are running studies on which colleges in which cities show the most delinquencies in student loans and how to adjust for the increased risk. They also use analytics to determine where ATM branches should be positioned and how much cash should be placed in them.

Q: Can you provide a use case of analytics in the banking industry?

Ans: Customers are increasingly using bank debit cards, and with every swipe, they create critical digital information. Analyzing usage patterns on larger data sets will not only reveal her/his buying preferences, but will also highlight engagements with the bank's affiliated merchants. With these insights, a bank's product-development and partnerships teams are better poised to decide if they should enhance partnerships with existing merchants or go for merchants with new and innovative products. They can also leverage the insights to decide on locations where a particular offer from a merchant may grab more eyeballs and mindshare.

A number of predictive models are used in the Banking Industry:

Category affinity: Predicts what product or service the customer will be most interested in.
Churn: Predicts whether customers are likely to renew their home insurance policy.
Credit card default: Predicts whether or not customers are likely to default on their credit card debt.
Customer segmentation: Segments customers into groups with similar demand characteristics, such as young educated and middle income, affluent and middle aged.

Sequence analysis: Predicts the kind of recommendation to make to each customer, based on what they have purchased. For example, a customer who obtains a mortgage is likely to want to purchase home insurance. A customer who has purchased travel insurance might want to activate a credit card for international use.

- **Determine category affinity in banking**
  You can determine customer affinity towards product lines by understanding the customer demographic information, purchase information, and browsing information.

  The Category Affinity model uses customer transaction data as input (transaction date, merchant, category, and price) and predicts the customer's category affinity. It uses a logistic regression model. Logistic regression classifies records based on values of input fields. The output is a predictive score that determines the probability that the customer will buy a product.

- **Predict churn in the banking industry**
  The churn model in the Banking sample predicts whether the customer is likely to renew their home insurance policy.

  The churn model takes in several variables into account. For example, the customer's monthly premium, the number of months since the policy was started, the number of claims that are denied, the number of months for renewal, marital status, income, and age.

- **Predict whether customers will default on credit card debt**
  The Credit Card Default model predicts whether or not customers are likely to default on credit card debt. The model uses customer information, such as age, education, number of years employed, income, address, credit card debt, other debt, and whether the customer has defaulted in the past.

- **Define customer segments in the banking industry**
  When you define customer segments, you profile the customers into groups that have similar demand characteristics.
  The inputs into the model are customer information, such as age, education, years employed, address, income, and debt to income ratio. The model uses clustering to divide the customers into various segments, such as young educated and middle income, or affluent and middle aged.

- **Sequence analysis**
  Sequence analysis predicts the kind of recommendation to make to each customer, based on what the customer has purchased. For example, a customer who obtains a mortgage is likely to want to purchase home insurance. A customer who has purchased travel insurance might want to activate a credit card for international use.

  Sequence analysis uses a Sequence model, which discovers patterns in sequential or time-oriented data. The model detects frequent sequences and creates a generated model node that can be used to make predictions.

7. OTHER APPLICATIONS

Q. What is Market Basket Analysis?

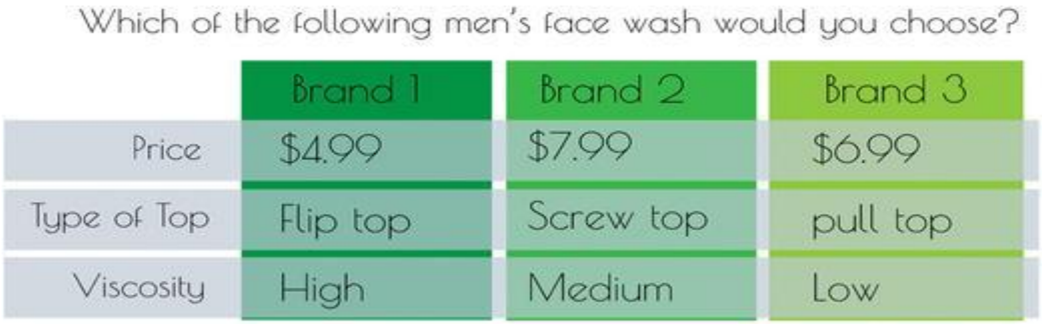
Ans: The term market basket analysis in the retail business refers to research that provides the retailer with information to understand the purchase behavior of a buyer. This information will enable the retailer to understand the buyer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new buyers (much like the cross-selling concept). An early illustrative example for this was when one super market chain discovered in its analysis that customers that bought diapers often bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically.

Q: What is conjoint analysis?

Ans: Conjoint analysis is widely used in market research to identify customers' preferences for various attributes that make up a product. The attributes can be various features, such as size, color, usability, price, etc.

Using conjoint (trade-off) analysis, brand managers can identify which features customers would trade for a certain price point. Thus, it is a much-used technique in new product design or for formulating pricing strategies.

Customers undergo a carefully designed survey, which showcases products with different attributes, and customers are then asked the likelihood of their purchasing the products. As shown in Figure 7-1, rather than asking their preference up front, the survey showcases different product-price combinations, to determine the latent needs of customers in context with the right price point.



Which of the following men's face wash would you choose?

	Brand 1	Brand 2	Brand 3
Price	$4.99	$7.99	$6.99
Type of Top	Flip top	Screw top	pull top
Viscosity	High	Medium	Low

A typical survey designed for conjoint analysis

Q: What are the three main steps involved in executing a conjoint analysis?

Ans: First, the product has to be divided into attributes and features.

The second step involves how these attributes have to be presented to the survey respondents. This step also includes deciding which rating methodology should be used for these attributes to be picked by respondents. The design decision takes into account factors such as number of respondents, time available for each response, complexity of product, and its features.

The third step is the statistical algorithm itself. The part-worth model is one of the simplest models available to assess the utility of each attribute.

Q: What are some of the ways in which an HR department can use analytics?

Ans: Human resource (HR) analytics forms the crux of the human resource (HR) function in most organizations and is a complex application of data mining and statistical techniques applied to employee-related data. HR analytics is a crucial technique for quantitatively measuring the outcome of employee engagement programs, and it provides corporations with the power to measure year-on-year contrasts with regard to numerous factors.

HR analytics provides organizations with powerful insights to help efficiently manage employees to increase productivity. There are various challenges in this field, including identifying the right data, capturing, storing and processing that data, and building the right model to increase return on investment (ROI) of the analytics function.

HR analytics finds its application in various business functions. The core functions of HR, such as recruitment and training, mergers and acquisitions, designing compensation structure, and improving performance appraisal processes are revolutionized by applying analytics to the historical data.
Some emerging niche areas in HR analytics are

- Employee sentiment analysis
- Predictive attrition models
- Net promoter score

Q: What are some of the specific questions that HR analytics helps to answer?

Ans: Some questions that HR analytics helps to find answers to include the following:

- Identifies the motive behind high employee attrition in an organization
- Identifies the reasons for lagging performance
- Finds strategies to improve team efficiency
- Measures the gaps in employee skills and identifies ways to fill them
- Identifies efficiencies in employee orientation (or comparable policies)
- Forecasts who is right to assume a specific role
- Makes the correlation between performance rating and employee performance

Q: What, briefly, is employee sentiment analysis?

Ans: Sentiment analysis involves extracting meaning and insights from large amounts of structured and unstructured data available to HR via various sources. These can be both internal and external sources. Internal sources include annual surveys, internal blogs and e-mails, etc. External sources include social media channels, blogs, and external surveys.

The data should be continuously tracked, analyzed, and scrutinized on key topics, which can be used to understand how employees feel about the company and what can be changed to make it a better place to work. Glassdoor.com, LinkedIn, etc., are employee-related portals continuously being used by various companies, to run sentiment analysis.

Q: What, in detail, is the predictive attrition model?

Ans: Employee attrition is predictable under stable circumstances, wherein a set pattern can be deduced from certain parameters influencing the employee and the organization at all times. Some of these parameters could be foreseeable, such as retirement age, or unforeseeable, such as company performance, external funding, management shake-up, etc.

However, who is going to leave, when, and why can be answered, based on analytical models developed as a result of data analysis.

Through predictive algorithms, companies' gain better understanding and can undertake preventive measures to counter employee attrition.

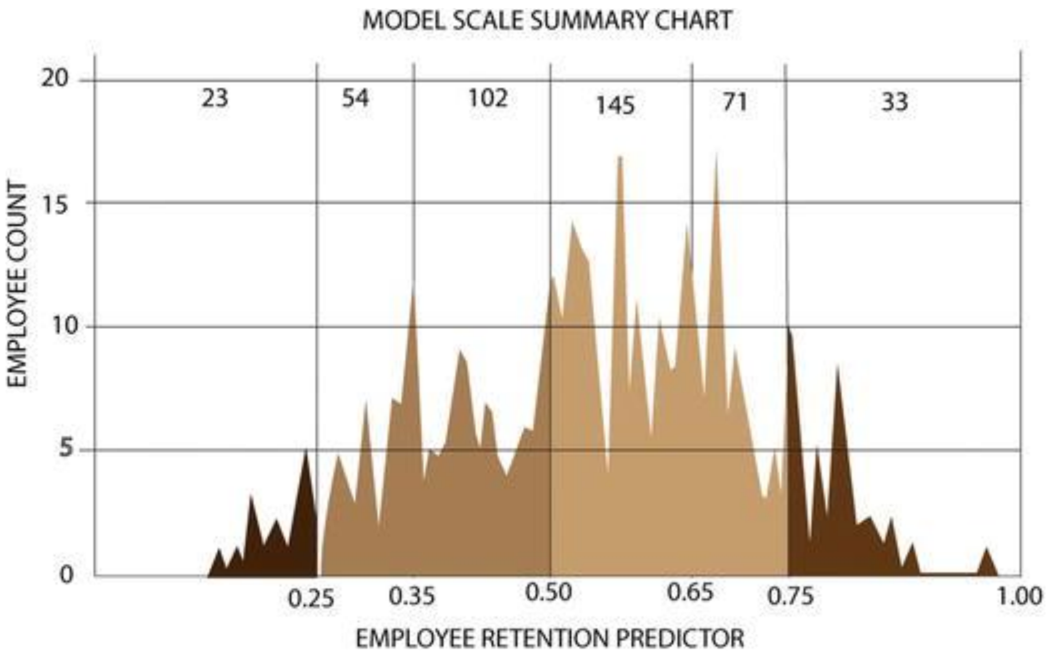
On a basic level, the model works by clustering/classifying employee profiles, based on various attributes, such as age, sex, marital status, education level, work experience, distance from hometown, etc., and generates various levels of risk of attrition. Occasionally, other parameters, such as performance over the years, pay raise, work batch, and educational institution, are also taken into consideration.
However, the accuracy of the model is directly proportional to the selection of parameters, which, in turn, leads to the generation of the "type" of predictive model most suitable for the organization.
A predictive attrition model helps not only in taking preventive measures but also in making better hiring decisions. Deriving trends in a candidate's performance from past data is important for predicting future trends, as well as to gauge new employees. Moreover, HR can use the employee data to predict attrition and the possible reasons behind it, and can take appropriate measures to prevent it.

We live in a data-driven world. From something as trivial as weather updates to GPS navigation, data is constantly being generated every second, in every field, which leaves us to decide how to turn it around to our advantage in our respective domains.

Q: How would you create a predictive attrition model? What, briefly, is the statistics part of this model?
Ans: Various statistical and machine-learning algorithms are designed to construct predictive models. For instance, classification models catalog employees based on their risk of leaving a company, whereas nonlinear regression models determine the probability of attrition when outcomes are dichotomous. Likewise, a decision tree model evaluates loss based on such factors as Gini index, information gain, and variation reduction. For models involving multiple parameters, decision trees tend to become very large and complex.

In such circumstances, a random forest method combines several decision trees, using multiple algorithms to classify and understand complexities and predictions. In addition, these models aim to provide good predictability. However, seamless implementation depends on choosing the right model.
Thus, different models are chosen, based on the aforementioned parameters, data availability, budget, computational power, and the requirements of decision makers. For example, in one organization, a model using an artificial neural network may provide better predictability than a decision tree model, but a decision tree model may be easier to understand and implement at a lower cost.
Thus, depending on the organizational contexts, different models have to be tried and evaluated before making the final selection.

Q: What would a typical output of a predictive attrition model look like?
Ans: The output depends on the chosen model. For instance, a logistic model produces scorecards for employees, based on their predicted attrition risk parameters, while the classification model catalogs employees into wider parameters, such as more likely or less likely to quit, high risk or low risk, etc.
Figure 7-2 gives the output of a typical logistic regression model, in which employees are scored on a retention predictor and charted on a distribution graph.



Output of a predictive attrition model
However, the bottom line is to keep it simple enough for HR managers to understand and implement. Changing the various factors helps in assessing the impact of changes and making the right decisions.

Q: What are some of the benefits of a predictive attrition model?

Ans: This model is helpful in doing the following:

- Evaluating employee requirements, strengths, and weaknesses
- Minimizing the cost of new talent acquisition, based on employee profiling and company requirements
- Analyzing and assessing loss in expertise and skill sets
- Measuring financial and productivity loss due to attrition
- Enabling planning and minimizing loss
- Providing a good understanding of workforce supply and demand
- Enabling the preparation of contingency plans, based on the insight and foresight provided by the model

Q: How can HR managers use analytics to promote employee effectiveness?

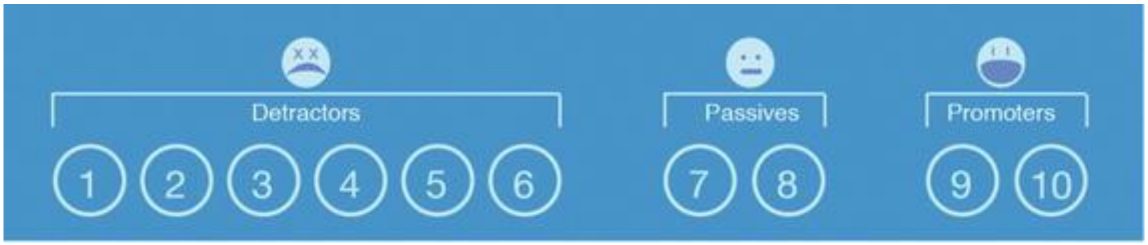
Ans: We can use a variable reduction technique such as factor analysis to identify the right attributes that will have the most impact on employee productivity. So, rather than focusing on many aspects of employee engagement, we can run statistical algorithms to identify attributes that have maximum impact on employee performance.

This approach can also be used for tasks such as improving skills, leadership hiring, employee movement, leadership identification, etc. HR managers can identify the attributes that affect skills at various levels and positions, using statistical algorithms, and can then extrapolate these to identify appropriate candidates.

Q: Briefly, what is Net Promoter Score, and how is analytics used with it?

Ans: The Net Promoter Score (NPS®) is becoming the standard metric for measuring customer satisfaction. NPS is a loyalty metric developed by Fred Reichheld, a Fellow of Bain & Co. and a board member of Satmetrics.

NPS is calculated by using the answer to a single question, evaluated on a 1–10 scale: "How likely is it that you would recommend [brand] to a friend or colleague?" This is called the Net Promoter Score question. As shown in Figure 7-3, people who award a score of 9 or 10 are considered Promoters of a brand, whereas any score of less than 7 is considered to come from Detractors of a brand.



Promoters, Passives, and Detractors in the NPS survey

NPS Score = % of Promoters - % of Detractors

In addition to the NPS question, each respondent can rate a brand/company across a number of service-delivery attributes.

Thus, statistical analysis helps to derive that, for each one-point improvement in these attributes, NPS will improve by xx points.

Variables that affect NPS can be assessed by using multiple regression techniques and beta coefficients, according to the following methodology:

1. Relationships are identified by correlating the variables.

2. Stepwise regression evaluates the different beta values for different variables, to identify how much a variable affects NPS.

3. The best beta values, the service-delivery level attributes, and other variables affecting NPS the most are identified.

4. The analysis explains xx% of the variance in NPS, i.e., for each unit improvement in the determined elements, the NPS score is improved by zz points.

5. Those attributes that are most important for all customers are identified at a corporate level, and this provides further scope for improvement.

Q: Briefly, how can analytics be used in the hospitality industry?

Ans: In the hospitality industry, analytics can be utilized to expand business operations, optimize marketing strategies, and increase occupancy rates and yield. For example, through analytics

- It allows a concierge to know which local tours are best suited to a guest's preferences, based on his/her past behavior.
- It lets a restaurant forecast the menu entries that are most expected to be ordered, based on what the weather is expected to be for the day.

- It helps in pricing decisions, such as, given the occupancy rate, what should be the right pricing of the room(s).
- It helps to optimize marketing strategies, such as what offers and campaigns should be sent to whom and when.
- It helps hoteliers cut down their energy costs without sacrificing guest comfort.

Q: Can you describe a use case of social media analytics

Ans: Social media has recently emerged as a goldmine of consumer-related information that supplements traditional ways of collecting information. Customers freely talk about their preferences and what they are interested in.

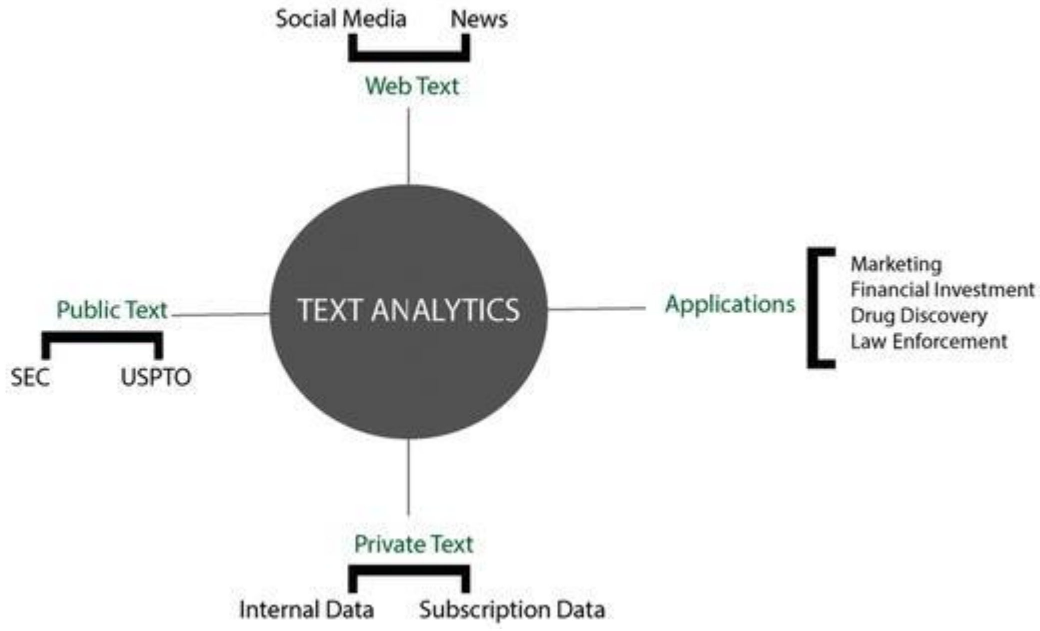
This data may include affinity toward a genre of music or a specific news website. Social affinity data can be a powerful way to help understand how brands relate to people's many interests, such as musicians, books, websites, movies, or celebrities.

This information is highly useful for brands, as it unearths deep-rooted preferences among customers, which might otherwise be difficult to identify.

By mining thousands of conversations and engagements over social media channels, marketers can identify key affinities for their brands. This can also be in the form of interest segments that align with your brand or with your competitors'.

Q: What is understood by text analytics?

Ans: Text analytics is the process of deriving high-quality information from free-flowing text. Organizations today have much more unstructured text than they realize. This may be in the form of comments from sales teams and agents, minutes of meetings, internal chats, e-mails, blogs, and much more. This text stores a vast amount of valuable information that most of the time goes unnoticed. Figure 7-4 details the text analytics inputs such as public and web text, as well as private text, such as a company's internal data.



Typical text analytics process
Using advanced natural language processing tools, we can derive insights from the text. For example, companies can mine thousands of tweets, to understand the customer sentiment for a brand.

Q: What typical process would you undertake for text analytics?

Ans: Text analytics involves a three-step process:

1. **Text mining:** More often than not, text might not be as easily available as imagined. Large amounts of unstructured data reside on various portals, blogs, handwritten material, etc. Text mining refers to the process of gathering this data, using algorithmic means.

2. **Text parsing:** Text parsing refers to the process of converting unstructured text into a form that is more readily analyzable. This may include writing algorithms simply to cherry-pick information from a text. Or, alternatively, to clean up the text (removing punctuation, numbers, stop words, or parts of speech tagging), to polish it into a well-rounded, analyzable raw data.

3. **Text analysis:** This is where the crux of the whole process resides—extracting those interesting nuggets that are insightful for the business. Techniques in natural language processing comes in handy, including pattern recognition, word frequency distribution, sentiment analysis (polarity), latent semantic analysis, word classification and categorization, single value decomposition, and word correlation.

Q: What is a word cloud? How is it used?

Ans: A word cloud is a visual representation of words in a text, with greater prominence given to words with higher frequency of appearance. Word clouds, because of their visual nature, give instant insight about which words are more important in the context of a text, as compared to others.
It is also possible to categorize words in some manner by color-coding them separately. In Figure 7-5, see a word cloud from the Wikipedia page on analytics.

analytics
data
analysis
Marketing
information
Business
jump
risk
digital
use
example
modeling
predictive
Retrieved
software
Wikipedia
customer
decisions
techniques
industry
insurance
governments
Creative
wealthy
science
many
fraud
Online
research
Big
Search
Optimization
Statistics
Privacy
page
marketers
events
may
Help
Learning
Leadership
organization

Word cloud

Q: Can you provide some key use cases of geospatial analytics?

Ans: Following are three key use cases:

LOCATION COMPARISON BASED ON POPULATED AREA DENSITY

Currently, finding the location for your next store, ATM, real estate asset, warehouse, etc., is more or less a market research activity. It's imperative that the right location for your next asset make a huge impact on your top line.

This is where geospatial analytics can help, providing a better estimate of the population around an area. With advanced remote-sensing techniques on geospatial data, we can estimate the population density around a particular location. This, combined with mapping competitors and other important sales generators (establishments that help increasing sales directly/indirectly), can provide additional information about important POIs (points of interest) for the most strategic location.

STRATEGIC LOCATION IDENTIFICATION

Companies are faced with resource crunches when deploying resources at higher concentrations of activity. Essentially, a brand's customer or key activity is spread throughout an area, and it's nearly impossible to cater to all these activities through the limited resources at hand. This leads to a strategic exercise into location identification. This may include identifying location for road shows or deploying personnel who maximize impact to customers or activities.

Geospatial analytics, coupled with data from other sources, can be extremely helpful in identifying these hot spots. The addresses of existing customers can be geocoded on the map, and areas that overlap the most within a one- to two-mile radius of these customers' locations can be identified. To advance it further, historic data such as time and geographic features can be coupled with the preceding and be fed into a machine-learning algorithm, to predict the probability of events.

AREA GROWTH IDENTIFICATION

One key decision parameter for many executives is how a location has changed over a period of time. This may include the number of housing facilities that have come up in the area, the change in cultivated land cover, and new roads and highways. This information can be used to identify areas with growth potential for real estate developers, or for retailers, in deciding where to plan their next store.
High-resolution satellite imagery of land areas can be extracted from different time periods, and the images analyzed using geospatial analytics. This gives the percentage change in land cover or in new housing in the area.

Q: How would you define big data?

Ans: With both computational power and storage increasing at reduced cost, the amount of data being generated today is profound. Most of that data is either unstructured or semi-structured.
With the advent of this large amount of unstructured data, the traditional method of storing and processing it are inefficient. Big data is large amounts of unstructured data that holds huge potential and value from being mined and stored but, due to its size, could not be stored on traditional database systems.

Q: What are the four v's that define big data?

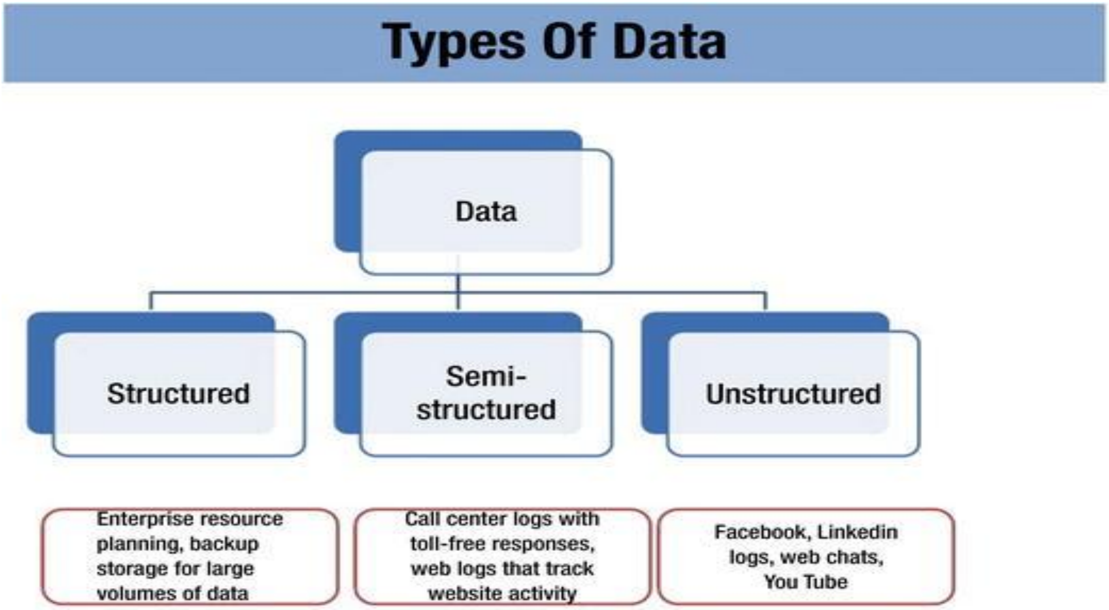
Ans:

- Volume: The amount of data has to be large, in petabytes, not just gigabytes
- Velocity: The data has to be frequent, daily, or even real-time
- Variety: The data is typically (but not always) unstructured (like videos, tweets, chats)

- Veracity: Uncertainty of data.

Q: Can you differentiate between structured, semi-structured, and unstructured data?
Ans: Most traditional data is structured, i.e., it can be stored in well-defined rows and columns. Legacy transaction systems are an example of structured data: all transactions are stored in relational database management systems (RDBMSs), with each row representing one transaction and each column representing attributes of that transaction.

Semi-structured data is partially stored in a well-defined database structure. Think of an XML file, which stores data but is not as well-defined as a database table.
Unstructured data cannot be categorized as structured or semi-structured and do not have a well-defined structure associated with how it is stored. Think of most tweets or blog posts. They contain relevant information to be mined, but this information is not structured. Special techniques are applied to extract this information.


Types Of Data
Data
Structured
Semi-structured
Unstructured
Enterprise resource planning, backup storage for large volumes of data
Call center logs with toll-free responses, web logs that track website activity
Facebook, Linkedin logs, web chats, You Tube

Market Research

Q. What is the process of market research?

Ans: Problemdefinition->Negotiationonbudgetwithclient->Qualitativeresearch(interviewbrainstorming)->IdentifyingparametersandbuildingQuestionnaireifprimaryresearch/identifyingsourceofdataifsecondaryresearch->Identifyingpopulationanddesignofsample->Preparinganalysisplan->datacollection->coding->verification &validation->Analysis of data->Interpretation->Presentation

Q. Do you come across Open ended; close ended; semi open ended coding? Assigning a different numerical value to parameters of estimate is coding.
Ans: No statistical tool can analyze data without numbers. (String/character/statement :-> numerical value.)

Q. What are the skip patterns in questionnaire?
Ans: If one question is yes then skip some question and answer some set of question. Like if you are using brand A you will answer some set of question if Brand B then some other set of question.

Q. What all major analysis performed in Retail market research?
Ans: Brand shifting, brand tracking, brand positioning, price analysis, marketing mix modeling, customer segmentation, opportunity for new product identification, Concept & product test.

Q. What is Market Basket Analysis?
Ans: The term market basket analysis in the retail business refer store search that provides the retailer within formation to understand the purchase behavior of a buyer. This information will enable the retailor to understand the buyer's needs and rewrite the store's layout accordingly, develop cross-promotional programs, or even capture new buyers (much like the cross-selling concept). A nearly illustrative example for this was when one supermarket chain discovered in its analysis that customers that bought diapers of ten bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically.

Q. I want to know for my product what my target population is. How do I do that using which statistical technique?
Ans: Need to do market segmentation using Cluster analysis. First do hierarchical clustering to identify no of segment of the market for similar product and then do K-means cluster to identify the levels of the parameters of the cluster to target.

Q. Why do we do Factor analysis?
Ans: To reduce the number of variables for a better presentation of the key factors.

Q. Usually how the analysis happens in a customer engagement analysis etc... Whether we do factor analysis first or cluster analysis or regression?
Ans: First we do factor analysis to reduce variables then with the factors we do regression to identify important factors then we do Cluster analysis based on the important factors.

Q. Why do we need weighting in market research?
Ans: To project characteristic with a sample for the population we have many constrains such as availability of proper sample. Like in population male female ration is 50:50. But the sample shoes 80:20 then we need to put a weight to make it in same proportion.

Q. What are the benefits of online research?
Ans: Online research offers greater access and a substantial time and cost savings over phone, in-person, and mail data collection methods. You can also more quickly receive results with online research.

Q. Will the quality of graphics accurately represent product concept, product, or advertising?
Ans: With the assistance of several major clients who extensively use graphics in their online research, SPSS Inc. has spent the past two years developing "best-in-class" graphic capabilities. We continue to invest in R&D in this area to ensure we maintain our high-quality standards.
A caveat: graphics quality largely depends on the original image and the end-user's equipment.

Q. My survey is very long. Can respondent stake a break and then return to complete the survey at a later time?
Ans: Yes. Respondents can logoff of a survey, log in again, and return to the survey at the exact page where they left off. This enables respondents to finish the survey at their leisure and helps improve completion rates.

Q. What kind of experience does your staff have with survey research and online surveys?
Ans: Our staff has numerous years of professional research experience from both the client and supplier perspectives. Staff members have experience programming a wide range of online survey projects, including:

- Concept / brand testing
- Conjoint studies
- Profiling and segmentation
- Website evaluations
- Customer satisfaction
- Brand image, awareness, and usage
- Commercial / TV program testing
- Print ad testing
- Multimedia evaluations
- Tracking studies
- Forecasting
- Business-to-consumer/
- business-to-business studies

Q. Can you program any advanced statistical techniques that require a dynamic/fluid survey experience?
Ans: Yes. We currently offer discrete choice modeling, a type of conjoint analysis.

Risk Analytics: Job Roles and Interview Questions

Job Roles:

- Building credit risk scorecard for screening new applicants (application scorecard)
- Developing behavioral scorecard for measuring risk level of existing customers
- Building PD, LGD and EAD Models
- Model validation and documentations
- Analyzing credit risk strategies
- Monitoring credit risk scorecards
- Forecasting portfolio losses and insights for loss provisioning
- Management and regulatory reporting
- Involve in risk identification, measurement, mitigation, and management
- Stress Testing and Scenario Analysis

If credit risk analyst role is for Business Banking then there may slightly different aspects and terminologies.
Similar to any other analyst job role, a credit risk analyst needs to have follow set of skills

- Domain Knowledge
- Technology & Tool Knowledge
- Quantitative Skills – Statistics, Machine Learning and Mathematical Skills

Job interview for a credit risk analyst role will cover questions across these three skills. Communication, leadership and other soft skills are relevant to a credit risk analyst as well.
Some of the domain specific concepts and questions in a job interview will cover

- Banking products
- Regulations and Regulatory Frameworks e.g. Basel II/III etc also country specific details
- Credit Decision Process
- Credit Decision System

Some of the commonly asked questions for Credit Risk Analyst job interview are
Domain Specific

- What are the differences between Mortgage, Personal Loan and Credit Card products?
- What are the key changes happened from Basel I to Basel II?
- What are the 3 pillars of Basel II?
- What is Vintage Analysis?

- What to you understand by reject inference? Why do you need? What are the commonly used methods are for reject inference?
- What is score alignment? How do you correct?
- What is the difference between bankcard, Private label credit cards and co-brand cards?
- What are the revenue and cost drivers of a credit card?
- What is the difference between transactor, revolver and surfer?
- How would you go about building an acquisition strategy for approving/declining credit card application?
- How would you build a credit line assignment/optimization strategy for a credit card?

Analytics and Data Science Related

- What are the commonly used analytical techniques for Credit Risk Scorecard development?
- What is rolling, performance and observation window?
- How do you convert probability values to scores?
- What is Points to Odds?
- How do you measure performance of a credit risk scorecard? What the model performance statistics?
- Explain model monitoring framework for Credit Risk Models?
- What is PSI and CSI?