SSPL™

# Predictive Analysis

## Better way to manage business

sspl.net.in

# Predictive Modeling

## Get Better Insight Faster

**Predictive modeling** is the process by which a model is created or chosen to try to best predict the probability of an outcome. Predictive Analysis is to predict future behavior of an individual on certain parameters based on past trends, also known as **Data Mining**. In many cases the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data. **Probability** is a measure or estimation of how likely it is that something will happen or that a statement is true. Probabilities are given a value between 0 (0% chance or *will not happen*) and 1 (100% chance or *will happen*). The higher the degree of probability, the more likely the event is to happen, or, in a longer series of samples, the greater the number of times such event is expected to happen. These concepts have been given an axiomatic mathematical derivation in probability theory, which is used widely. In Insurance we can use the **Predictive Modeling** to calculate if not 100% correct, but close upto 80% accuracy of the behavior a customer on the certain pre-defined parameters. Before arriving at Predictive Model of scoring, a detailed data analysis is required to look for the trends on various parameters. Based on the analytic inputs, scoring can be defined and implemented in the model. Insurance companies see this new normal firsthand, providing data networks and technology convergence to bring information and access to their customers in the moment. As a result, it's become critical for insurance to have the same type of fast access and interaction with their customer data and analytics. The current competitive landscape demands it. Analytics has become the backbone of business strategy through its ability to obtain valuable information from customer interactions and behavior, and then convert it into customer loyalty and revenue uplift. Analytical modeling is widely used in the insurance industry to tackle difficult tasks all to increase ROI while maintaining customer satisfaction. Another reason for increased adoption of predictive analytics is because of the increased level of adoption of data warehouse and BI/MIS reporting solutions in many insurance organizations, where such existing DW/BI capabilities can easily be extended to adopt more effective predictive analytical solutions. The recent increase in availability of large data (Big Data) from a variety of data sources has opened a new dimension in predictive analytics space, enabling the insurers to take advantage of such large source of data, along with it existing traditional data, to gain useful insights that and reap commercial, business and IT benefits to the insurers and to their customers.

## Insights to Action

Companies that systematically apply predictive analytics to operational decisions, especially those pertaining to customers, outperform their competitors. The best practitioners create advantage within multiple core functions such as originations, marketing, account management, customer retention and collections by applying predictive analytics in operations.

## Large Complicated Data Challenges

In insurance organizations, historical data pertaining to policies, claims, customers, partners and market data plays a significant role in building highly effective predictive analytical models such as policy risk scoring models, claim fraud risk scoring models, subrogation models and customer churn models. Organizations often maintain historical transaction data (claim, policy, customer data) at least up to 10 years and such structured data, along with other unstructured content, can be of enormously large size, in the order of multiple terabytes. Processing such large quantities of data during data mining, can pose some of the following challenges:

**Enormous data preparation effort** – To apply correct statistical data mining methods such as linear, nonlinear, log-linear, classification/clustering (k-factor), regression, neural networking including decision trees etc the data needs to meet structure as required by the chosen model. Data preparation effort would usually include the following, to make the target data to be fit to model with the chosen model, to improve the accuracy of the mining.

**Data type conversions and transformations** – Multiple data mining models require dependent and independent variables to be numeric or categorical (in some cases). Such data type conversions can take enormous effort with large data sets.

**Missing value analysis** – Missing or incorrect values in the data can directly impact the data mining results. Before applying the data mining models, it is necessary to perform missing value analysis using techniques such as missing data pattern identification, estimate missing value statistics such as means, deviations, correlations (pair wise variables) and fill the missing values with estimated values using regression, estimation and multiple imputation methods. Large data sets can take enormous time and effort to perform missing value analysis.

**Higher levels of cleansing and noise reduction effort** – Higher the volume of the data, larger will be volume of the incorrect or missing data (noise), requiring expensive cleansing and noise rectification processes.

**Target data set (Universe Population) preparation** – Large volumes of data (Big Data sets) can impose new constrains when defining training (development) data sets and testing data sets, that can truly represent the underlying population universe. Also to improve the accuracy of the prediction, it is important to run the predictive algorithms against the complete universe of the target data, in which case, the extremely large data sets can induce performance issues.

**Performance of the mining algorithms** – Using traditional data mining tools and platforms, large volumes of data can impact the performance of mining processes, taking considerably higher time for the execution of mining algorithms, almost making it impossible to process against large production data sets. The performance can become even more important with non-linear and mixed models, which use multiple iterations to improve the statistical estimates in the models.

## The Role of Technology

IT plays a critical role in adopting practically feasible predictive analysis solutions in an insurance organization. The role of IT is evident in all stages of predictive analytical processes. The following are the key areas, where role of IT is highly crucial in adopting effective predictive analytical practices:

- IT plays a critical role to facilitate tools, software package and solutions for data mining and predictive analytics that can be easily be leveraged by business users.
- IT can play a big role is providing solutions and infrastructure, which can process very large quantities of unstructured data, emerging from new sources such as web click streams, twitter, and internet blogs, that can provide very useful information relating to various aspects of the business.
- Data quality plays a critical role in determining the accuracy and effectiveness of the predictive analytical models. IT must a play a strategic role in maintaining the data quality in down-stream and upstream systems.
- Enterprise level predictive analytical processes are highly dependent upon the detailed and summarized high quality data held in purpose-built data warehouses. IT can play a big role in deploying high performance, scalable data warehouse/BI and next generation predictive analytical platforms, embedded analytics, distributed data warehousing and hardware based ETL solutions, that can support real time or near real time predictive analytics against segmented target population or against full target population, comprising of big data with structured and unstructured data formats.
- IT has to leverage to automated tools, solution frameworks and accelerators to automate the predictive analytical processes.
- IT's role also extends in provisioning highly secured data warehouse and predictive analytical platforms that can ensure full compliance to data privacy and associated regulatory requirements, while dealing with both internal and external data sources (Big Data sources).
- IT can play role in reducing the overall TCO of BI and predictive analytical platform, by providing innovative solutions including emerging cloud based predictive solutions.
- IT's role is critical in facilitating business continuity and scalability of the BI/Predictive analytical platforms, that are critical to business operations in insurance.
- IT can be instrumental in promoting and familiarizing the data mining and predictive analytical practices across the organization, by deploying user friendly and business friendly predictive analytical tools and solutions.
- In the future, predictive analytics moves further towards handling more and more complexity in the models, and IT can play a major role in facilitating the predictive analytics, using models with higher degree of complexity.

## Simply the Complication

The complexity in the model increases with the number of the independent variables (factored variables/agents) and their non-linear relationships, with predicted variables (claim risk score, policy score etc.) and co variables. Usually, higher the number of variables, the larger will be the complexity of processing the predictions. For instance, propensity for claim fraud can depend up multiple factors such as age, gender, location, country, insurance product, past history etc. The complexity in predictive analytical models can be simplified by carefully analyzing the models and applying various dimension reduction techniques, to simplify the models, before taking them forward for further analysis. Some of the following types of dimension reduction techniques are helpful to simplify the models:

**Factor analysis** – Techniques can be used to analyze and screen the variables, and choose the right set of variables that can explain higher degree of relationship with the predicted variables and thus simplifying the complexity in the model.

**Correspondence analysis** – To analyze the relationships between the variables to select the right set of variables for further analysis.

**Scaling techniques / multi-dimensional visualizations** - are used to determine (visualize) the perceptual relationships between the objects and thus to select useful variables, that can be used further in a predictive analytical model.

Complexities associated with large data in training, validation data sets, can be addressed by using structured data sets (samples) that are representative of complex design requirements of the underlying model, so that the result outcomes are valid. Various complex data sampling techniques such as clustering, multi staging sampling, stratified sampling, non - random sampling, and unrestricted sampling etc. can be used, to carefully choose the training, validation data sets, to improve the validity of the results.