Caveats with Data Analytics

The combination of large data sets and observational data means that data analysis activities are often at risk of drawing misleading inferences. In this article we describe five of these dangers. Within the operations research and analytics community these problems are well understood. Unfortunately, the solution is not easy and often tempers down the enthusiasm, and to be cognizant that rather more complex models are necessary.

- Hidden nuances of data collection – Many a times the analyst is not aware of the process of data collection and the hurdles or bottlenecks of doing so. However a detailed understanding of the same renders the data analysis more holistic and improves handling observations that otherwise may lead to diametrically opposite inferences. Say for example we are studying the galactic observations from telescope to understand the distance of space objects from earth. One of the primary assumptions here will be that further the object is from earth, dimmer will it look under the telescopic lens. However there may be interstellar and intergalactic gas and dust clouds, which attenuate radiation. So it violates the primary assumption. Of course these phenomenon is known to scientists but things get more complicated or messy in situations where the data collection process are not so well understood, or, even worse, when the possibility of such events is not considered.

- Process change or Non-stationarity -Any inferences from data analysis is not valid if the there is any change in the underlying process or pattern of events. For example, a model of herbivorous animals browsing propensity on grass lands will be quite useless when the animal population declines rapidly. Any historical data on browsing will have little or no relationship to patterns after the population decline.

- Correlation Vs Causation – Its very critical to appreciate the fact that statistical correlation doesn't necessarily mean causation. High sale of umbrellas may have high correlation with number of car accidents but the cause can be either rains or snow fall. Similarly in a scatter plot we may have a positive correlation between age and height for a given range but unless we have the adequate process knowledge we would not be in a position to infer whether weight causes height to increase or vice versa !

- Pre-empting Vs Intervention – Many a times using analytics we like to prevent frauds. And to enable that we set certain thresholds either in transaction values or frequency or in absolute scale. For example in credit card transactions there are symptoms in terms of frequency and transaction values. In car insurance sector, timing of n accident claim is a good determinant for a better diligence to validate if the claim is fraudulent. However if these thresholds are some how themselves predictable then

there can be interventions to avoid those detection mechanism in doing fraudulent transactions.

- Data aging – Age of data is also very critical in analytics. Real estate mortgage repayment credit history was never so bad before the sub-prime mortgage crisis hit US. However it is not easy to understand the life cycle validity of data. For example defaults in loan repayment may only happen after a minimum number of installments. For auto industry this threshold can be minimum of one to two years. For real estate the threshold can be little longer. So one has to wait that much time in the industry to understand the behavior pattern of users. However even if a strata of specific profile of consumers start behaving in a particular way after a age of 3 years since beginning of a process, there may be enough reasons that in three years the new consumers of the profile may have changed their behavior.