**Week 9**

**Article #1: An introduction to data mining and other techniques for advanced analytics**

What are the key differences between statistical analysis and data mining? (page 140)

* Statistical analysis is limited to samples that represent a population, vs data mining has access to an entire target population’s data.
* Data mining requires different approaches to traditional statistical methods, which may be misleading or overfit the model and produce unhelpful results
* Data mining datasets have huge numbers of variables that can’t all be predictor attributes, so data mining solutions usually provide tools for selecting relevant attributes for analysis
* Statistical analysis aims at identifying a statistically significant model for prediction, but there’s no guarantee this model will perform sufficiently or be of business value, while data mining includes diagnostic results to indicate likely business benefits from a model.

Describe tools for advanced analytics (page 149-151)

·  Data visualization : transforms massive volumes of data into multidimensional pictures and animations that highlight areas of interest to help understand complex data. Examples: scatter plots, heat maps, maps.

·  Text mining : the discovery of previously unknown information or concepts from text files by an automatic extraction process (data mining). Examples, Attensity, Clarabridge, IBM (SPSS), KXEN, and SAS which use linguist analysis to extract facts from unstructured text.

·  Social network analysis : identifies groups of people connected in some way by applying network theory concepts such as “nodes” and “links.” Example, SNA can extract potentially useful new variables about the size, strength, and composition of customer’s calling circle for use in data mining projects.

·  Contact optimization : takes all of the analytical model predictions as inputs and searches for an optimal allocation of products and channels to customers over time. Furthermore, the allocation has to satisfy budget constraints, contact rules, and minimum/maximum volume limitations.

* Inbound optimization is primarily concerned with delivering the ‘best’ solution for each individual customer who contacts the company’s call centre or logs onto their website. This can be viewed as an extension of the customer management or CRM system — to supply the next best offer for each customer, based on a set of predicted propensities for the available products.
* Outbound contact optimization aims to find the ‘best’ solution at an individual level and at the same time meet overall outbound marketing business targets and constraints. It enables the business to forward plan the communication mix and estimate the return over a future period, as well as compare alternative communication strategies.

How do you mitigate the risks of data mining? (page 152)

(a) Data quality issues — the data being mined must be of high quality, consistency and integrity. Failure to achieve this can be critical, both at the modelling and deployment stages in the process.

(b)  Untrained users working with highly automated modelling tools can produce misleading or nonsensical results.

(c)  Producing mountains of unusable or non-actionable results. Being able to identify patterns in a data warehouse is only useful when there can be a business application. Having lots of patterns without profit-generating applications can be a costly distraction.

(d)  Poor evaluations of model efficiency, or lack of standards for evaluating descriptive results, can result in misuse of the findings and no gains from the process.

(e)  Certain technical requirements apply in the modelling stage (such as not extrapolating outside the domain of the data), which is again why users have to be fully trained.