Unit 9 Reading Questions

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

and early 20th centuries, before the advent of computer technology. Methods were required for making inferences based on relatively small samples drawn from the corresponding populations. **The theory was developed for testing hypotheses and measuring signiﬁ cance of results, taking sample size into account, since analysis at population level was not a viable possibility. At the same time, the number of records and the number of attributes for which measurements were recorded were sufﬁ ciently small to enable each variable to be examined individually and transformed as appropriate for analysis and modelling purposes. DM, on the other hand, is applied to databases that typically hold an entire population of customers, together with thousands of variables that summarize their transactional behaviour, payments history, campaign responses and so on. Therefore, in any project, the data miner is no longer restricted to working with small samples — the full customer base is available if desired. However, this requires some differences in approach from traditional statistical methods — statistical techniques may give misleading results if applied to a vast sample size, which carries risks of over-ﬁ tting the model or producing unhelpful results in whi**ch every variable appears to be statistically signiﬁ cant. Furthermore, in DM, the dataset is liable to contain a huge number of candidate predictor attributes (variables), for example, volumes and values of transactions by product, channel, brand or period — far too many to be individually assessed and transformed manually. DM solutions ideally provide automated tools for selecting relevant attributes and recoding them in the form of variables for use in analysis. A further key difference is that statistical analysis will aim to identify a model that is statistically signiﬁ cant — that is, outperforms a random prediction — based on a set of signiﬁ cant predictor variables. However, this provides no guarantee that the model will perform sufﬁ ciently well to be of business value. DM goes further than that, by including diagnostic results to indicate likely business beneﬁ ts from the model. The assessment is produced by using two methods in combination:

(a) Prior to modelling, a random subset of data is excluded from the analysis, for use in evaluating the power of the model developed on the remainder of the data — this excluded subset is known as a ‘ h old-out’ sample, and is more likely to give a fair indication of model performance than if the model development sample were used. (b) Various types of tables and charts are produced in order to assess the predictive power of the model, using the hold-out sample. For example, if a model has been built to predict campaign response, then the lift chart will show how response rate varies by the probabilities predicted from the model (often grouped into deciles). This will help the users to decide whether the model is likely to deliver enough beneﬁ t to justify its deployment and select the model deciles that should be targeted.

Lastly, having built and evaluated a DM model on a sample dataset, the model will be deployed by applying the scoring algorithm to all ‘ X ’ million records in the customer database. Therefore, facilities for large-scale model deployment are essential — the form that this takes will vary from package to package, as we will see below. Both DM and statistical analysis require that the data are organized as a simple rectangular table, where the rows (or records) represent individuals (eg customers) and the columns contain structured variables (eg demographics, usage or purchasing behaviour). Often, much effort is required in order to assemble this analytic dataset, as we discuss in the following section. The variables in this dataset are ‘ structured ’ , in the sense that each column contains either numeric or character (categorical) values coded in a consistent format. However, an increasing amount of information is captured nowadays in an unstructured form, for example customer comments, accident reports and e-mail requests. A technique known as ‘ text mining ’ may be used to read unstructured data and derive facts that can be represented by structured variables, and included in analytic datasets — this approach is discussed under ‘ Other Tools for Advanced Analytics ’ below.

        Describe tools for advanced analytics (page 149-151)

* Data visualization

Scatter plots, which enable interesting features to be discovered by overlaying other attributes onto a plot, combined with highlighting and interactive drill-down in order to ‘ d ig more deeply ’.

Heat maps, which are used to locate ‘ hot spots ’ across a graph formed by interlacing two variables.

Maps, which are used to examine data displayed geographically and search for possible relationships between the geographical objects. Maps may be relevant at any spatial scale — for example, national or regional maps can display the dispersion of a target market, while a map of a supermarket layout can display intensity of trafﬁ c or purchase patterns by department.

* Text mining

Text mining solutions typically use linguistic analysis to extract facts from unstructured text. A number of different text mining tools are available, from suppliers such as Attensity, Clarabridge, IBM (SPSS), KXEN and SAS.

* Social network analysis

SNA has been gaining traction over the past few years, as analytics users have been starting to learn that SNA metrics are correlated with customer loyalty. For example, in the mobile phone sector, SNA can identify the members of each group or ‘ c alling circle ’, determine the central communicator or ‘ key inﬂ uencer ’ and extract various metrics about the strength of relationship within the group. If the mobile operator is concerned with spreading marketing offers by word of mouth, then these key inﬂ uencers will be the best people to inform. Similarly, a good predictor of defection may be that a subscriber is in frequent contact with a person who themselves has recently defected.

* Contact optimization

answers this question is known as ‘ contact optimization ’ . Contact optimization ‘ sits above ’ DM, in the sense that it 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. The differences between inbound and outbound communications imply a need for separate contact optimization approaches. 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 ﬁ nd 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. Suppliers of contact optimization software include Experian, SAS, IBM (SPSS), TCP Marketing Solutions and Unica.

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

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 proﬁ t-generating applications can be a costly distraction. (d) Poor evaluations of model efﬁ ciency, or lack of standards for evaluating descriptive results, can result in misuse of the ﬁ ndings and no gains from the process. ( e) C ertain 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.