Analytical Modelling Success

Barry Leventhal offers seven tips for successful analytical modelling projects.

I was recently gloating over the chapter on “top data mining mistakes”, in a data mining textbook, when it made me think - what would my personal top tips be for successful analytical modelling projects? So here are my thoughts.

1. Plan and improve your analytical process

Process planning and improvement might seem ‘nice to have’, however they become essential if your main analytics output will be a targeting or segmentation model that needs to be applied to records in a database. The methods of specifying and implementing the model algorithm should be carefully defined, in order to automate the process as far as possible and avoid time-consuming stages. ‘Best in class’ users employ modelling software that outputs SQL code (or equivalents) generated to perform the scoring calculations in-database, but it’s surprising how many others still struggle with model scoring. This implies that those users are forced to limit themselves to a small number of models that are only occasionally re-scored – thus forgoing most of the benefits that targeting and segmentation should be helping to deliver.

2. Consider segmenting your target variable

In many modelling projects, the analyst is able to choose – or at least recommend – the target variable that will work best for the data and problem on hand. In such cases, it’s worthwhile giving consideration to the target variable that will yield the most accurate and useful predictive model. Some years ago, I was involved in trying to ‘rescue’ a site location model that had been developed for a chain of off-licences. A large amount of data had been assembled for each existing outlet, at site and catchment area level, and was used to develop a model to predict total sales. Unfortunately, this model appeared to have very poor predictive power, and was judged to be too weak to present to the client. On re-examining the data, I tried building separate models for the different components of total sales, such as wines, spirits and so on, and found that each of these models actually worked well. If this approach had been taken, it would have produced a useful predicted sales profile, as well as a good overall prediction.

3. Predict customer behaviour ‘by cause’

On a similar theme, it can sometimes prove beneficial to develop separate models that predict customer behaviour **by cause** rather than a single ‘overall’ behaviour model. A classic example occurs in the mortgage market - if you are building an attrition model to identify which customers are most likely to terminate their mortgages, in order to carry out some business retention activity, then it makes best sense to create separate models for ‘home movers’ vs. ‘re-mortgagers’. These groups are likely to differ demographically, as well as having very different reasons for ending their mortgages. And if the mortgage lender does not hold customer data on reasons for leaving, then it can be worthwhile carrying out a follow-up survey on recent leavers, in order to collect this information (for a sample of past customers) and use it to build the models.

4. Integrate your data sources

Most companies tend to store their data in separate places across their functions – billings data in the accounts department system, campaign history in the marketing database, usage records in the data warehouse and so on. When creating the analytic dataset, for analysing an event such as customer churn, it’s highly beneficial to link together the information from these disparate sources into a single customer record that contains all aspects of the customer relationship. The more complete picture can reveal previously unknown patterns, some of which may link to important outcomes such as churn. It can further help to overlay external data, including geodemographics, demographics and lifestyles, in order to understand the likely market-wide involvement of each customer. Analysis of churn in the mobile phone market is a good case in point – customers will churn for a variety of reasons, such as price, finance, network performance, and influence of friends/family and therefore a wide range of data sources will need to be leveraged in order to capture the predictors of churn.

5. Analyse your detailed data

Continuing on the theme of predictor data, the single richest information source is likely to be the detailed interactions with your customers – such as call detail records, financial transactions or shopping baskets. However, the data volumes are vast and therefore these sources tend to be pre-summarised at an aggregate level in standard monthly extract files. If you could access the **detailed** data – perhaps initially for a sample of customers and limited time period – then you'd be able to create more specific predictor variables, and test out your own hypotheses about behaviour patterns that might be causing the target outcome. You'd also be able to quantify the accuracy improvement in your model and the return that would be achieved if the newly discovered predictor variables could be included in the standard extract, and thus justify this development. From personal experience, the gains can be significant – taking this approach to build a survival model for mobile phone subscribers, the model that used detailed data predicted churn twice as accurately as the equivalent model based on aggregated data.

6. Use analytical techniques in combination

It's unlikely that application of one analytical technique alone will be enough to solve a complex business problem – just as a car mechanic would probably use more than one tool to repair an engine. A series of steps is more likely to be required, applying different analyses and passing the data through these processes – with checks to ensure that meaningful results are being produced at each stage. One of my all-time favourite projects was to create a set of small area estimates for consumer demand, using a combination of market research, census statistics and census micro-data - this required many models and estimation stages, but worked well because careful testing was carried out at each stage, building up to mapping the final set of estimates.

7. If it looks unusual, it’s probably wrong

Long ago, a famous media statistician coined Twyman’s Law, which says that “any figure that looks interesting or different is usually wrong” - I believe that Twyman’s Law still holds good today. It implies the need to investigate unusual results and seek out the reasons for their occurrence – this should involve quizzing people with deeper understanding of the business, in order to discover whether a logical explanation exists for the strange result that you’ve observed. Some years ago, a propensity model was built to target home loans within a database of bank customers. The most powerful predictor, by far, was found to be prior take-up of a personal loan – this fact was accepted by all, until the home loan manager explained that customers often took out a personal loan in tandem with their home loan, in order to build up the deposit for their purchase. Therefore, the target outcome had effectively ‘leaked back’ into the predictor data, thus making the model invalid.

Dr Barry Leventhal is director of independent London-based consultancy BarryAnalytics.
Tel: +44 (0)7803 231870
barry@barryanalytics.com