REAL WORLD DATA SCIENCE

Matthew Leitch extols the considerable business benefits of bespoke data analysis

Buzz phrases such as big data, machine learning, and data science conjure up images of the big tech companies like Google and Amazon doing complicated and expensive things with computers. While for others, artificial intelligence will one day put some of us out of a job. But in the decades before that happens on a large scale, if it ever does, there will be plenty of opportunities for intelligent, numerate people to get useful results from simpler initiatives involving some of these technologies.

Those initiatives could, for example, help clean up data efficiently, target efforts to boost sales, keep customers, reduce the costs of operational mistakes and help justify investment in decisions where the outcomes are otherwise too uncertain to support a major investment. The biggest advantage of data science methods over just slicing and dicing data is the ability it offers to quantify the combined significance of multiple variables at the same time.

But what do you need to know to run a simple, low-budget exercise to gain more from your data with data science methods?

RESOURCES

You could spend hundreds of thousands of pounds on statistical analysis tools and then far more on implementing a data warehouse to pull data from different databases into one database to feed those tools. If your ambitions are more modest, you could download the software used across the world in universities to do data science work, completely free of charge. That software is called R, and the core software can be expanded with literally hundreds of packages, usually written by academics, that are also free. R has support for a huge range of statistical regression methods, plus support vector machines (SVMs), random forests, neural networks, symbolic regression, and many others.

There are packages to allow you to load more types of data, and at least one package dedicated to really big data files, which otherwise are a problem for R. The R documentation is pretty comprehensive and standardised, though not the most user-friendly. There are also many books published to explain R and its packages. Often an author writes a book and a package together.

Besides being free, R has one more key advantage: it is the software most university students doing mathematics or statistics learn to use for data science work

Besides being great and free, R has one more key advantage: it is the software most university students doing mathematics or statistics learn to use for data science work. That means most new graduates in mathematics or statistics today have used R, have written programs in R, know about the most important packages, and are used to reading R documentation and working out what to do.

An alternative to employing someone in the UK is to go offshore, but there is an important reason for preferring someone who will be on site with you and speaks perfect English. It is hard to learn anything useful by data analysis alone. As explained below, a much more effective approach is to cycle between analysing data and talking to people. A good understanding of the organisation, or at least a good ability to ask people questions and understand their answers, is crucial and likely to be more important than a more advanced degree.

Realistically, this type of work is intellectually difficult and you need very clear-thinking, mathematically competent people to do it. It helps if they can also conduct a fact-finding meeting competently. Experience and knowledge of real organisations and systems is enormously valuable, but the more you demand the more you will have to pay to secure their services.

GENERAL GUIDELINES

It is vital to understand the types of discovery likely to come out of data science work, to anticipate them and to act on them to turn insights into benefits.

First, discoveries made are very unlikely to be statistical links that can be turned into more sales. More likely, they will be problems with data in your systems, especially non-financial systems that do not have the benefit of traditional accounting controls. The analyst might just see these as an irritation and try to work around them, but they might be more useful than that. Correcting those errors might translate into reduced workload in your organisation, fewer issues for customers and more reliable information. Imagine a pipeline of discoveries being turned into benefits and think about how you can log the items and make sure everything gets appropriate action.

Second, is to combine information graphics with statistical modelling. Very often the graphics come first and are part of checking for errors in the data. They are also vital for explaining findings and convincing people they make sense. Just running the numbers is a big but common mistake.

Third, cycle between data analysis and inquiries with people. Number crunching alone typically produces disappointing results. Human intelligence still has a crucial role. Imagine you have downloaded some operational data from a company system and found a correlation between two variables, A and B. It's probably a rather weak correlation, as they usually are in real data science work. What does it mean? Does A drive B? Does B drive A? Is there something else going on that drives both A and B? Why is there any connection at all? Why isn't it a stronger link? To find out more, select some specific items (for example, invoices, orders, or whatever is relevant to the data set) that are typical of items where the link appears. Then go to see people who worked on those items directly ask them what happened exactly, and then ask them if they can see any reason why there might have been that statistical link.

This method accesses information that may not even be on your computer systems. Human intelligence can also help identify the direction of causal links detected as correlations. For example, if a customer buys a 'happy birthday' banner and a birthday cake then we know what that means, but to learning algorithms it is just a correlation.

Finally, human understanding of the world, including events people remember but which were not recorded on a computer system in an easily identifiable way, can help us do 'feature engineering'. This is where the analyst constructs a new variable, or feature, by combining some others. For example, suppose you want to predict if someone will pay their next bill. You have a history of their past bill payments. It is human intelligence, typically, that allows you to construct a new variable showing the number of consecutive unpaid bills leading up to the new one. In practice, this will be a powerful predictor, but most machine learning algorithms today will not find themselves.

It is often the case that getting data is harder and more time consuming than analysing it, so another useful guideline is to try hard to analyse the data you have in many ways to explore possible explanations and raise more interesting questions. That is a lot easier when you have detailed suggestions from your inquiries with people.

THE NATURAL SEQUENCE

Data science work is complex, and sometimes progress depends on what you find, what co-operation is given, and other factors that are unpredictable. However, there is a natural sequence that should guide your planning.

The first wave of insights is likely to come from carefully scrutinising individual data sequences and pairs of them before trying to combine them in multivariable statistical modelling. This scrutiny will usually find errors in data, or at least anomalies that look like they probably are errors. Also, simply looking at distributions with charts can reveal unexpected things happening in the organisation.

The second wave of insights comes from using regression to make predictions. This enables you to focus effort efficiently on particular cases. For example, you might be able to predict when customers are likely to leave you and so target efforts to retain them. Such predictions are usually easy because they typically use data you already collect and require no changes to procedures or special arrangements with other teams.

Sometimes progress depends on what you find, what co-operation is given, and other factors that are unpredictable

However, this form of analysis does not reveal which factors are the most important drivers of results. That requires much more data than just making predictions and can be very difficult if your data has multicollinearity. This is where potential drivers correlate with each other, making it difficult to separate their effects and decide which of them is producing the result you are interested in.

Also, some learning methods do not show the contribution of each variable.

Neural networks and SVMs are examples of machine learning tools that are good at prediction, but terrible at finding drivers. You cannot examine the models created for meaningful clues to the key drivers, so the best you can do is look at what difference it makes to add variables to a model. In contrast, linear regression provides coefficients and significance levels that nicely summarise the importance of each variable used.

The third wave of insights happens when you can identify which factors are the most important drivers. This might require more data, more inquiries, the right algorithms and a bit of luck. There is no guarantee that the data you capture on your systems contains any powerful individual drivers of the results you are interested in. This is a major reason why, in practice, regressions tend to perform far below their theoretical capability.

The fourth wave of insights comes when you organise colleagues to conduct experiments rather than relying on mere statistical links, as in regression analysis. An experiment is a very specific research technique that involves randomised assignment to groups, each treated differently. This can compare the effect of the treatment with other treatments or with doing nothing special. No amount of regression can reliably quantify the value of an intervention, but an experiment can.

Persuading colleagues to be rigorous in carrying out a proper experiment may be easier if you can offer them efficient samples to work with. If you have an intervention that many people are already quite enthusiastic about and think will be effective then it is difficult to persuade people not to use it on every possible case. Conducting an experiment will require that at least some cases are left untreated so that a fair comparison can be made.

For example, suppose an organisation wants to get more customers to pay electronically. Based on just anecdotal evidence and gut feeling, most senior people are now enthusiastic about engaging a call centre to call customers and try to get them to change their payment methods. However, as this is an expensive option with highly uncertain results it might be worth doing an experiment to test exactly how much difference it makes. For that you need a call centre to call some customers, but not all of them. Some executives argue for just going ahead with all customers and not testing; others suggest only calling the customers most likely to be persuaded. Neither of these approaches will give you an unbiased experiment, so if you use them you will not know how much difference the calls have made.

Persuading colleagues to be rigorous in carrying out a proper experiment may be easier if you can offer them efficient samples to work with

One way to compromise is to devise a method of predicting which cases are most likely to benefit from the intervention and then divide the population into groups with varying priorities. With each group, divide cases between treatment and control groups, but not equally. Do treatment more often in the high priority groups, but make sure you still have enough to provide a comparison in all groups. When you present the final results, you should see that the intervention works better than doing nothing special, and the difference is greater with higher priority groups. This is more informative as well as biasing selection towards the commercially valuable cases in a way that usually makes the overall value of the exercise greater.

Another potential problem is that the people doing the intervention in your experiment may fail to process all the items in your treatment group. If they fall short, then your sampling is invalidated. To get around this, consider randomly dividing the population into several equally sized tranches and then divide each tranche by priority as described above. This way, every completed tranche gives you valid comparisons, and the more tranches get completed the better. If a tranche is not completed then all the data from it has to be left out.

The final comparison of treated and untreated cases does not require the complex machine learning tools that are common with regression, but regression models (and good charts) can be used to squeeze more learning from your sample.

CONCLUSION

Data science need not be a huge investment for many organisations, but some intelligence and knowledge of the technical basics is needed. Practices like combining good information graphics with algorithms, can make a big difference to productivity. By working through a natural sequence that starts with checking your data and moves on from there, you can create a pipeline of discoveries and turn them into benefits. ●