Datazin

MORE ABOUT PREDICTION...

by Kees Schippers - Independent Database Marketer

MORE ABOUT PREDICTION...

"My physician needs to know where my ***headache*** *stems from so she can decide on proper treatment or further tests to run."*

"The weatherman has to know something about ***tomorrow's weather*** *or we'd have noticed wouldn't we?"*

"I should like to know ***tomorrow's stock prices*** *before buying any today."*

The marketer wants to know different peoples' preferences to decide which product to offer to whom.

Well..I can't simply guess tomorrow's stock prizes.

The physician cannot just instantly know what causes my headache.

Same goes for the weatherman, the marketer, or anyone else who needs information that is not directly available.

So, how should one deal with that?
The answer is: if you can't directly know what you want to know, infer it from what you do know.

If you can't directly know what you want to know, infer it from what you do know.

How I predicted stock prices

I haven't the slightest knowledge about the stock market. But what I - and most people like me - do know is that stock prices change over time. And it is precisely that knowledge that I will try to convert to money.

Historical stock price data are digitally available. So I can easily calculate average day prices. For every day price I can calculate an index to the previous day's price or the day before that. I can also do this by week or month or even by the hour or minute. In short, all sorts of trend calculations can be made over the recent or remote past at any moment in time, which we can call 'Moment Now'.

Since I have historical data I have many 'Moments Now' for many different stocks. To this 'Moment Now' knowledge I can then attach stock price information after that moment using the same historical database. Then it is a matter of statistical analysis to see if there is a relationship, a 'correlation', between what I know at 'Moment Now' and what happens after that moment, call it 'Moment Future'. If there is a statistical relationship, then that relationship can be used to predictfuture stock prices, at least to a certain extent.

Such a relationship would enable me to estimate what will happen with different stock prices later. The goal is to buy those stocks that - on average - will have a higher price later. I was able to predict which stocks rose on average about 1% within a few days based on recent changes in volume and price while the market generally fell a bit. That is, I was able to identify the specific subgroup that rose in that period. Not all stocks of this specific subgroup went up of course, but the average result over these stocks was 1% increase in price. It was the timing of buying and selling that made the difference. Alas, keeping, buying and selling stocks also costs money, a sort of handling fee. That in itself vapourises my dreamed of profits. Further, the predictive power wore off pretty fast. Nevertheless, not bad for a couple of days work of data knitting and analysis. Taking this to the next level certainly seems a good idea. Especially when combined with specialists knowledge.

Furthermore, a good examination of the details of these relationships - how are they precisely related - can yield valuable information for the development of new hypotheses which can subsequently be tested.

Taking this to the next level certainly seems a good idea. Especially when combined with specialists knowledge.

My **physician** might ask me all kinds of things: about my past, about my behaviour, about the symptoms. These facts I can tell her, although describing symptoms is not always easy. She then will use my answers to create a hypothesis, which can be further tested. Certain symptoms are known to be related to specific causes. The strength of this relationship - and her skills as a physician - determine the certainty with which she can indicate the cause of my misfortune.

Likewise the **weatherman** must rely on what he knows now to predict what will happen later. And the **marketer** uses accessible information to infer unkown information.

Although the prediction in my example on stock prices might not be very impressive, commercially it is quite interesting.

Some illnesses are easily diagnosed, others are more difficult.

We can all check the accuracy of weather predictions for ourselves. I am quite impressed by the percentage of correct weather predictions made, even many days ahead.

And for the marketer, you can check your mail and e-mail on commercial offers and decide how much you think is relevant for you.

The general method

So what happens conceptually in all these examples is the relation between known facts – I will call these 'predictors' - and the unknown facts – I will call these 'outcomes' - can be established by looking at historical data in which predictors and outcomes are both known. That relation is then generalized to new cases with known predictors but as yet unknown outcomes.

Sounds simple and it is. Thousands of researchers in hundreds of universities and other research institutes do this for their daily work. Scientists usually are more interested in causes rather than prediction. They therefore prefer to manipulate what they think causes an event in a structured way and consequently measure the effects: experiments.

Scientists usually are more interested in causes rather than prediction.

Many things cannot be experimented upon, leaving us to deal with data we just find in the world as it is. In their book Freakonomics Steven D. Levitt and Stephen J. Dubner provide some great examples on this latter type of research, often called 'field research'. I mention this book because it radiates enthusiasm for the type of data exploring I like so much. It is the first non-technical book on analysis that can actually be read by people who do not have a strong statistical background. Actually, it is the second. The first is How to lie with statistics by Darrell Huff. Because of the brilliant content, as well as his fluent style of writing.

Do not forget

Two technical, very important things still must be mentioned briefly. First, make sure enough data are analysed to be

reliable. "What is reliable" and "when is there enough data" are quite difficult statistical matters. For the statistically trained this technical article may help answer such questions. There is no specific rule of thumb on what is enough data, but there is one simple overall statistical rule: the smaller the sample, the less trustworthy the results. Drawing different sample sizes from a known population and calculating means and percentages from those samples will provide a feel of how reliable samples actually are.

Second, you can generalise your findings only to the population the sample was drawn from. That is a **validity** issue. Is the sample representative to suit your research objectives? That could simply mean that if you gathered your data only on weekends, for example, you could not then assume your findings are valid for weekdays. The general message is to always think critically about possible limitations on the type of data used for analysis.

Kees Schippers
Independent Database Marketer

www.datazin.nl