5 Common Mistakes People Make in the Name of Statistical Analysis

White Paper

5 Common Mistakes People Make in the Name of Statistical Analysis

White Paper

Decmber, 2011
www.mu-sigma.com

5 Common Mistakes People Make in the Name of Statistical Analysis

Imagine you are a regional sales head for a major retailer in U.S. and you want to know what drives sales in your top performing stores. Your research team comes back with a revealing insight - the most significant predictor in their model is the average number of cars present in stores' parking lots.

Or perhaps you've sat in a business operations meeting where hours were spent explaining a two percent deviation in weekly sales even though the random variation in the data is five percent. Maybe you've sat in quarterly Marcom budget planning sessions where you are using a sophisticated market-mix model to allocate your spend but with a "small" caveat – this quarter, company leadership has decided to increase brand spend 10-fold and decrease non-brand spend proportionately.

These situations epitomize some of the most common mistakes in statistical analysis. In the work of performing analytics services across multiple verticals and geographies, common themes arise:

1. Sophistication in statistics compensates for lack of data and/or business understanding. Increased understanding and acceptance of sophisticated statistical techniques in business has resulted in enhanced availability of packaged solutions. These solutions have twin advantages in increasing the usage of statistical tools by business users and reducing the lead time required to gain results/insights. However, the convenience has lead to an added temptation of supplementing lack of data/business understanding with sophisticated statistics. This has resulted in overreliance on algorithmic approaches to analytics problem solving.

In this approach, business understanding is used to validate the outcome of analytics -- not necessarily the analytics process. A common symptom of this problem is the prevalence of esoteric modeling and data mining techniques without enough inquiry in to their appropriateness and applicability for the problem at hand. Unfortunately, it is model accuracy that often becomes the final arbiter. This results in a classic trap of choosing the technique that gives maximum accuracy over the one that makes most business sense. This can be best avoided by striking a balance between algorithmic and heuristic approaches, which is essentially a balance between a highly accurate model and a model that makes business sense. Tilting to either extreme is dangerous.

2. Extracting meaning out of randomness. Any data that you encounter has non-zero "meaningful pattern" to "noise" ratio and the art is in being able to isolate and explain the meaningful pattern and being able to tolerate the unexplained noise as error. But this is easier said than done. Suppose you are relying on a sales-forecasting model to help validate the quarterly targets you received from finance. Wouldn't you want your model to be as accurate as possible so that you can set realistic goals? This is a reasonable expectation, but overemphasis on model accuracy might land you in uncharted waters. Ask any statistician or an expert modeler and you will hear that statistics provides enough tools and implements to make the data say whatever you want to hear.

A common outcome of this problem is that you can get a model that is able to explain minute variations in the data that it modeled but fails miserably on any new data. This is call called the "Problem of Overfitting" in statistics. This happens because noise is a random phenomenon that is beyond the control of your business or even known

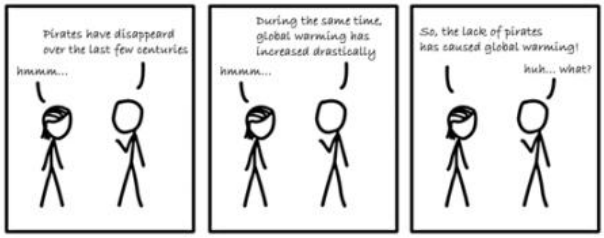
external factors (that is why it is noise!), and the part of your model that tries to explain this noise fails when it is trying to look at new data.

The way to control for this problem is to always compare in-sample validation (model accuracy on the data over which it is built) with out-sample validation (model accuracy on the data which model has not seen). In data mining parlance, these are called training and test data sets, respectively. For the model to be stable, meaning explaining only systematic patterns, the in-sample error should be reasonably close to the out-sample error. What is reasonable depends on the particular context of the problem, but typically if difference between the two errors is greater than 10 percent than you have reason to worry.

3. Correlation versus causation – modeling will help uncover causal relationships. This problem can be illustrated with two simple examples. Suppose you are a meteorologist who has a poor track record of predicting rain. Your model has the usual predictors such as weather, temperature, extent of cloud formation, wind speed, etc., but still is not reliable. It is clear that you are missing some significant predictor. One day, your analyst comes to you with a breakthrough: a variable that is a very significant predictor in the model and is able to validate historical data very accurately. Unfortunately the variable is weekly sales of umbrellas. This example shows that modeling does not establish direction of causality; in this case we all know that rains cause sales of umbrellas and not vice versa, but there is no way for the model to know this.

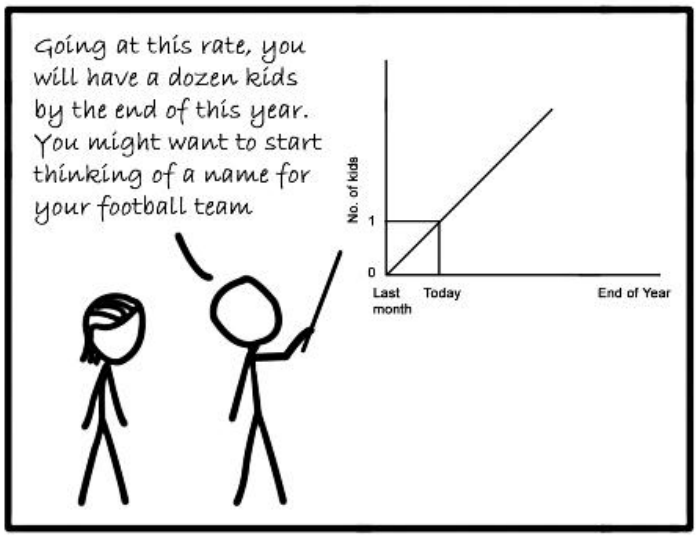
Another example: Suppose you are a regional sales head of a leading home appliance manufacturer and you want to figure out what drives sales of air conditioners in a region. Your insights team comes back to you with an excellent driver model that is able to explain the sales of air conditioners fairly accurately. The only problem is that the most significant variable in the model is the sale of aerated drinks. This example shows that modeling does not correct for the presence of confounding factors.

We know that sale of aerated drinks is not causing the sale of air conditioners and vice versa; ambient temperature is the common factor that impacts both. Because the impact is directionally the same, they are highly correlated and the driver model will naturally show this variable as a significant predictor. In this case, sale of aerated drinks is a confounding variable that should not be present in the model. The recommended way to overcome this problem is to start with a hypothesis matrix. A hypothesis matrix lays down hypotheses connecting every predictor with the predicted and also records the direction of impact. For example, in the above problem, a hypothesis connecting price with sales will read, “As price increases, the sales of air conditioners drop.” No predictor should go in to the model unless there is a well thought-out business hypothesis.



4. Extrapolating the models way beyond the permissible limits. Well-designed statistical models can answer a lot of business questions, but one has to realize that there is no perfect model that is free from all constraints. Judicious use of statistical models can aid business decision-making, but not being aware of a model's limits can be counterproductive. A statistical model is based on underlying data and is subject to the limitations of the data captured. For instance, a model to predict sales cannot take into account the impact of an earthquake on sales if the historical data has never captured earthquakes. Hence, this model will not be able to predict sales accurately in the event of an earthquake. These "Black Swan" events are often the cause of considerable distress; the subprime crisis being the most recent example.

Another example is marketing mix models that are used to assess the impact of marketing on sales. These models are often used as tools for scenario planning where the business user aims to estimate the sales based on different spend scenarios. It is important to realize that any model is only accurate in the range of data it has seen and, therefore, if the scenario is drastically different from history, there is a very high chance of error. For example, if the marketing division decides to increase spend by 5x, the same model might not be as accurate as it has been developed based on the historical spend. To understand intuitively, any model is just an interpreter that interprets data into a language we can understand. If the data does not speak about earthquakes or high marketing spends, the model will not be able to interpret it accurately.



5. Imputing missing values with mean or median is the best way of treating missing values. Any real life data being used for statistical analysis is likely to have quality issues and missing values in variables is one of the most recurring issues. Therefore, it becomes imperative for an analyst/statistician to impute missing values to avoid loss of data and retain maximum information. Often we encounter scenarios where there are 5-10 percent missing values in variables and we are inclined to impute them with the mean or median value. While it does the job in certain cases, extreme caution needs to be taken before imputing missing values as it might have significant consequences on model behavior and interpretation of parameters estimates.

It is important to realize that missing values can tell a story and help us better understand the business dynamics in many cases. Hence, it is necessary to look deeper whenever a variable has missing values before coming up with an imputation. For example, while conducting an analysis on premiums for a large health insurer, it was observed that 5 to 6 percent of values were missing. Further analysis revealed that the missing values were only for one state in the U.S. for a certain time period. Research revealed that the company was temporarily banned from operating in that state due to a legal issue. It is, therefore, recommended to look for the cause of the missing values before jumping into imputation.

This is by no means an exhaustive list but is certainly a representative one of the types of errors encountered in application of statistics to business. Some of these mistakes stem from incomplete understanding of statistics, some from the incomplete understanding of underlying business and the rest from the inability to marry the two together. With the advent of data analytics and decision sciences, our decisions are being increasingly impacted by these errors, which can result in major implications for our business and therefore the need for the business executives to appreciate, sense and avoid these common pitfalls.

Shashank Kumar Dubey is senior manager at Mu Sigma, and has vast experience in analytics consulting with multiple Fortune 500 clients. His experience spans across multiple industries - retail, airlines, hospitality, insurance, technology and telecom across multiple geographies. He works closely with the client teams and business executives in creating, operationalizing and driving the consumption of analytics.

Dhiraj Rajaram is the founder and CEO of Mu Sigma, an analytics services company that helps clients institutionalize data-driven decision-making. Dhiraj is responsible for the vision, strategic direction and leadership in Mu Sigma. Before Mu Sigma, he advised senior executives across a variety of verticals as a strategy and operations consultant at Booz Allen Hamilton and PricewaterhouseCoopers.