Shirley Coleman's comments on:

Experiences with Big Data: Accounts from a Data Scientist's Perspective

20th March 2020

The authors have produced an excellent essay with many interesting insights based on their broad experience of data analytics. I would first like to respond to their comment about people saying: "Here is the data, do something". Absolving responsibility for data is not always a bad thing. Sometimes it is sensible and worthwhile for the data scientist to do Exploratory Data Analysis and offer insight. In our paper relating to automotive aftersales (Smith et al., 2019) we note that the interaction between data owner and data scientist can be a mixture of push and pull where the data owner requests certain analyses thereby pulling information from the data scientist (often motivated by customer requirements) but is also open to having insight pushed back to her/him.

I would also like to defend the occasional use of fast and simple solutions, as such a response is sometimes just what is needed. For example, if you are tasked with advising a company where to test for pipeline integrity across 3000 Km of tundra, what do you do? You can attempt to construct a semi-variogram based on changing soil types and you can apply domain knowledge such as knowing that there is a higher likelihood that the pipeline may fail at curves, joins or changes in soil substrate or that 16 % of failures are due to ground movement. But in the end, you have to make a recommendation, well aware that it is not likely to be the best solution, but at least it is a solution and will allow some action to be taken.

One emerging issue is the fashion for companies to employ a token or trophy data scientist, to make their company look good and to be seen to be keeping up with their competitors. This practice can lead to a dangerous watering down of data science skill sets as the company may employ a less skilled person because they do not know what a data scientist actually does (Vicario et al., 2019). Even if such a company does hire a highly skilled person, unless the company has prepared an appropriate work plan and made supporting resources available then that person may be under-utilised and likely to leave for a more fulfilling role. Only a data scientist, or a person who is at least data-aware, can determine how data science can help the business and identify who to employ as a data scientist.

A similar issue arises with companies buying equipment to look good or keep up appearances. For example, it is not uncommon for companies to buy data analytical tools that no-one ever uses. I recall going to a healthcare provider and being told that they were fully able to understand their data and extract insight from it because they have a new computer package. When asked how well the package functions, it became apparent that although the director who sanctioned the expensive purchase thought the equipment was used, no-one had worked out how to use it or had actually turned it on. In another example, minute-by-minute fuel consumption data from high tech sensors is valued by ship's personnel as an early warning device where sudden fluctuations can indicate potentially catastrophic changes to the engines. Ship's staff contentedly assume that the data is fully scrutinised by admin staff and analysed on their behalf to calculate efficiency and performance but in reality this is highly unlikely and the data most likely goes unnoticed.

Data scientists should be encouraged to think widely and creatively about a problem. If they can't find any patterns explaining an issue, they should take a little extra time to see if anything else of value can be extracted from the data. This sort of open question should be included in training courses to encourage data empathy and deeper understanding of the value of data. For example, a dataset of questionnaire responses can be provided along with the hypotheses that the data

collection aimed to investigate. The students are asked to test the hypotheses but then to interrogate the data further to see if there are any important patterns that could point to hypotheses that could be examined in future work.

In general data should be used as raw as possible. I recall issues of uncertainty whether to assess product quality from a process industry on the basis of manufacturing plant output or on the basis of company output; plant output shows the performance of the process whereas company output is the sold product after quality has been adjusted. Analysis of each set of data has its own merits. Plant data shows input material consistency and equipment performance whereas company output shows what customers are buying and what is passed on into the supply chain. Similarly, we have the issue of confusion arising from data that is typically adjusted as supporting meta-data is received. For example, gas consumption can be raw or be post error adjustment once known meter errors are allowed for. The result is multiple datasets which are more or less similar but include adjustments made after different periods of time. All these datasets contain important information but care has to be taken that everyone knows which version is being used.

Analysing process data can lead to insight but the analysis is more satisfying and can be used for prediction if end results are known. For example, a company has developed a bespoke expert system and charges people to use it to obtain recommendations of products and services that would be useful to them (Ahlemeyer-Stubbe et al., 2018). The company knows the challenges that users need help with, and their personal circumstances, and the products and services offered to them but does not know whether the product or service was actually purchased or not. Furthermore, purchase behaviour alone is not enough. For example, in a retail business the data analyst has to make sure that returns are identified to balance against sales so that they know what products were actually retained by the customer and can take account of this information in the analysis.

****

References

Ahlemeyer-Stubbe, A, Coleman, SY. (2018) Monetising data – how to uplift your business, Wiley, Chichester; ISBN: 978-1-119-12513-6.

https://onlinelibrary.wiley.com/doi/book/10.1002/9781119125167

Smith, W, Coleman, S, Bacardit, J, Coxon, S. (2019) Insight from Data Analytics with an Automotive Aftermarket SME. Quality and Reliability Engineering International, 35(5), pp 1396-1407.

https://onlinelibrary.wiley.com/doi/full/10.1002/qre.2529

Vicario, G, Coleman, S. (2019) A Review of Data Science in Business and Industry and a Future View. Applied Stochastic Models in Business and Industry, pp 1-13.

https://onlinelibrary.wiley.com/doi/full/10.1002/asmb.2488

Biography

Dr Shirley Coleman is Technical Director of the Industrial Statistics Research Unit, School of Mathematics, Statistics and Physics, Newcastle University and a visiting scholar at the Faculty of Economics, Ljubljana University, Slovenia. She works on data analytics in Small and Medium Enterprises (SMEs) and contributed a highly ranked impact case study to Newcastle University's Research Excellence Framework. She is the academic lead on Innovate UK funded Knowledge

Transfer Partnerships developing data science capabilities, and specialises in statistical and machine learning techniques applied to company data. She publishes in trade and academic journals and is co-editor of several books. She is a past President of the European Network for Business and Industrial Statistics (ENBIS), an elected member of the International Statistics Institute and a Chartered Statistician of the Royal Statistical Society, instrumental in mentoring early career statisticians and developing relationships with business and industry.