The importance of domain knowledge for successful and robust predictive modelling

Received (in revised form): 9th February, 2021

Andrea Ahlemeyer-Stubbe
Director Strategic Analytics, servicepro GmbH, Germany

Andrea Ahlemeyer-Stubbe is Director Strategic Analytics at servicepro GmbH, and former President of the European Network of Business and Industrial Statistics. She is co-author (with Shirley Coleman) of 'Monetising Data — How to Uplift Your Business' and 'A Practical Guide to Data Mining for Business and Industry', as well as a frequent lecturer at various universities and speaker at industry conferences.

servicepro Agentur Agentur für Dialogmarketing und Verkaufsförderung GmbH, Trausnitzstraße 8, Munich 81671, Germany
E-mail: ahlemeyer@ahlemeyer-stubbe.de

Agnes Müller
Senior Analytical Consultant, servicepro GmbH, Germany

Agnes Müller is Senior Analytical Consultant at servicepro GmbH. She is involved in projects and workshops for customers both large and small from different industries. Combining technical perception with textual skills, her focus lies on making complex analytical cases and results understandable for customers and other people who are unfamiliar with analytics.

servicepro Agentur Agentur für Dialogmarketing und Verkaufsförderung GmbH, Trausnitzstraße 8, Munich 81671, Germany
E-mail: mueller@ahlemeyer-stubbe.de

Abstract Domain knowledge helps to build more precise and robust predictive models and thus obtain better insights. In the course of preparatory work, it helps inform what questions to ask, define the key fields to examine more closely, and identify where and how the insights from the analysis can support business goals. As this paper will discuss, it is also of great benefit when it comes to selecting or reducing variables, supplementing missing data, handling outliers or applying specific binning techniques. This paper argues that data scientists cannot rely on technical knowledge alone; rather, they must acquire relevant domain knowledge and familiarise themselves with pertinent rules of thumb. The paper also highlights the importance of maintaining close contact with the people who collect and prepare the data.

KEYWORDS: predictive modelling, domain knowledge, binning, dummy variables, data preparation, missing data, data mining

INTRODUCTION

The 4th industrial revolution is on its way. Processes in the economy and society are becoming increasingly data-driven and automated. But in the search for efficient solutions, knowledge regarding business and cultural contexts must not be neglected. Even if it is difficult to express in data sets, domain knowledge is still indispensable.

Industry is on the verge of new working conditions and processes. The changes that have been triggered by the avalanche of data

Figure 1: Predictive modelling utilises data from the past to predict future probabilities

are just as disruptive as the changes that took place when electricity replaced coal and steam as the main source of energy.[1]

Much like how electricity caused a huge leap forward, companies can now benefit even more from the digital revolution. They can use insights from data in all areas of their business: to improve their production, communication, strategy, pricing and purchasing. It is therefore important to rethink the way analytics is done and how data is collected and used.

Simply implementing a tool to run predictive modelling is not enough. The assumption that the algorithm will find the best solution on its own is flawed. The tool will simply sit on the Big Data solution deployed and crawl through the data. To obtain valuable predictions, it is important to be very thoughtful. Predictive modelling is about much more than just using the right algorithm — it is also important to know under what circumstances the information will be collected and the predictive models or other analytical insights will be applied.

In a nutshell, predictive modelling can be likened to prejudice: just as people form prejudices based on incomplete, historical and sometimes false information, a predictive model learns from historical but mostly incomplete data (Figure 1). To overcome this weakness, predictive modelling needs as much information as possible.

Predictive modelling belongs to the class of analyses also termed 'supervised learning'. Supervised learning is used to estimate an unknown dependency from known input and output data.

Input variables might include quantities of different articles bought by a particular customer, the date the purchase was made, the location and the price paid.

Output variables might include an indication of whether or not the customer responds to a sales campaign. Output variables are also known as targets in data mining.

In the supervised environment, sample input variables are passed through a learning system and the subsequent output from the learning system is compared with the output from the sample. In other words, the goal is to predict who will respond to a sales campaign.

The difference between the learning system output and the sample output can be thought of as an error signal. Error signals are used to adjust the learning system. This process is done many times with the data from the sample, and the learning system is adjusted until the output meets a minimal error threshold.

The process is not unlike tuning a new piano. Such fine-tuning may be done by an

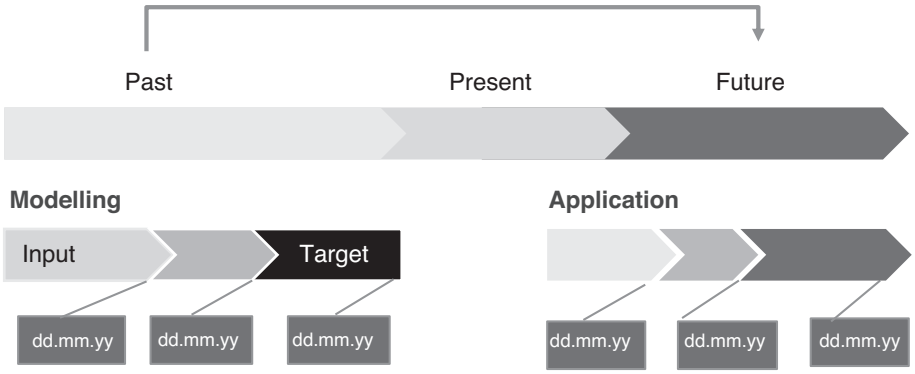

Figure 2: Predictive modelling — Timeline training and application

expert or by using an electronic instrument. The expert provides notes for the training sample and the new piano is the learning system. The tuning is perfected when the vibrations from the piano keys match the vibrations in the expert's ear.[2]

Supervised learning produces a model, and the value of the model depends on how well it either explains or predicts the patterns in the dataset. The process of conducting an analysis can often be immensely valuable, simply because of the focus on collecting a clean, reliable set of data. Sometimes the benefit of the model can be demonstrated in terms of expected financial savings or increased profits.

The quality of the model can be checked by applying the model to a new sample of data and comparing predicted target outcomes with observed target outcomes.

Another way to check the validity of the model is to compare the results with what is already known about the data and business structures behind said data. One should always look at the results from a purely business point of view and check that they are reasonable.[2]

So, the general idea is to use past information (data) to predict the likelihood that a certain target will be reached.

To train this kind of model requires historical data (note that the last click is also a past click) — data that may be considered complete and that may be divided into input and output (target) data.

Input data may be defined as data recorded before the output (Figure 2). It is essential for this information to be available when the predictive model will be applied (in future).

Training the predictive model to understand the relationship between the input data and the output (target) requires an algorithm.

Common modelling techniques for solving this type of prediction problem include linear and logistic regression models, decision trees or random forests and neural networks — the merits of which are discussed elsewhere. Of greater interest to the present paper are the impact and importance of data and how to add domain knowledge to existing data (although the importance of following a well-established data-mining process (Figure 3) in order to generate successful and robust predictive models is worth reiterating).

To be successful, people and processes working with data should consider more than just the available data and suitable methods of transformation. They should consider that there is and will be a tremendous change in the way that data are generated and used in the future.

One of the pitfalls of data related to process, marketing or sales is that the data is collected under certain circumstances

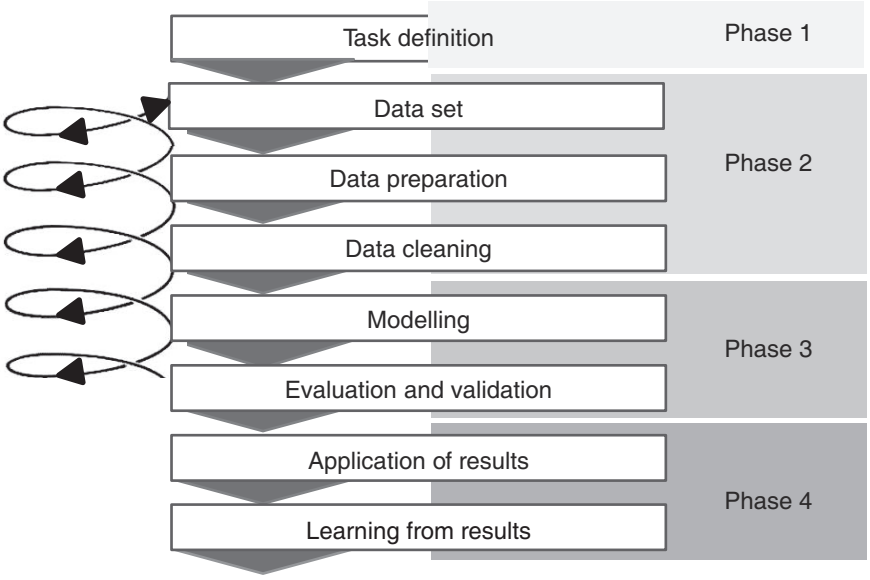

Figure 3: The data-mining process

and that tracing an item back for analytical purposes is sometimes not possible.

Indeed, the data used in analytics and stored in data warehouses or Big Data solutions tend to have been collected simply to answer a specific question or to control a specific part of the production cycle or fulfil particular business intelligence needs.

By contrast, the data used to solve complex problems as part of a designed experiment will be highly controlled. This is the kind of data that data scientists and statisticians are used to working with. It is not, however, the kind of data available for most predictive modelling projects.

DATA IN PREDICTIVE MODELLING PROJECTS

Data gathered from websites, sales, lead generation, points of sale, workshops or production, or stored to feed control needs or make it possible to track back in the event of a certain problem arising, is not the same as data collected via a designed experiment. Table 1 provides an overview of the major differences.

It is therefore important for data scientists to bear in mind that most of the data they use will have been collected opportunistically.[3] Where nothing has happened, nothing is stored. This consideration should also include the market environment and other external factors that could influence the outcome or have an impact on production or customer behaviour. Such factors may include weather conditions, or the fact that a competitor is celebrating an anniversary, or simply the different cultural background of different sales or production areas. This kind of information is seldom controlled, stored and audited, and the data are usually not to be found in the organisation's Big Data solution or data warehouse.

HOW DOMAIN KNOWLEDGE POWERS DATA PREPARATION

To illustrate how influential domain knowledge can be, consider the following analogy. As most people intuitively know, demand for ice creams is higher on hot, sunny days than on cold, wet ones (Figure 4). Now, although most data

Table 1: Differences in how data are collected

	Experimental	Opportunistic
Purpose	Research	Operational
Completeness	Full	Sparse
Size	Small	Massive
Hygiene	Clean	Dirty
State	Static	Dynamic
Market/environment	Controlled influence	High impact and uncontrolled

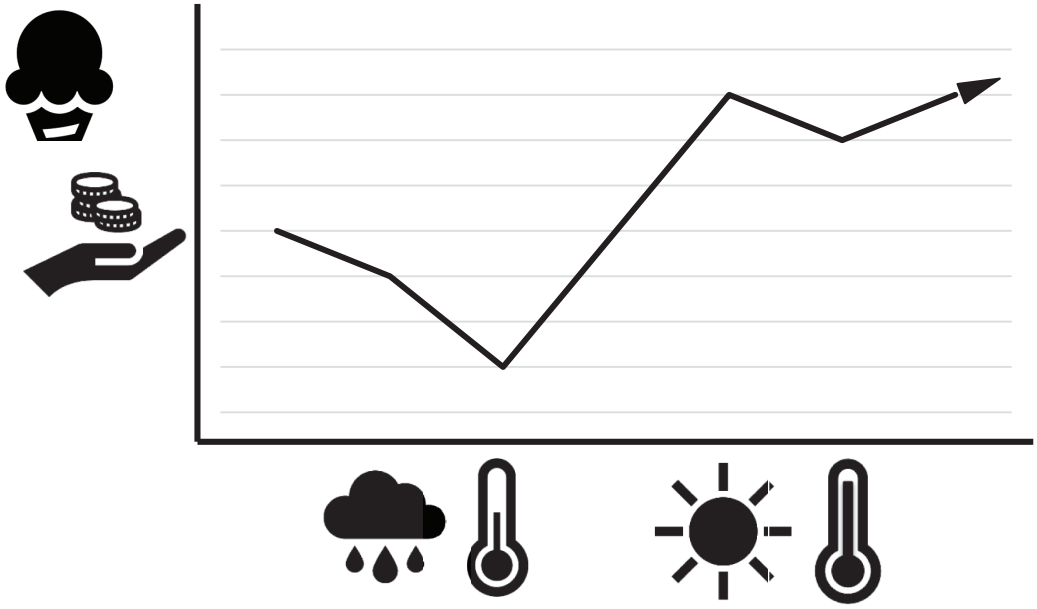

Figure 4: Domain knowledge — Correlation between weather and sales of ice cream

warehouses will record the types of products sold, the volume of those sales, and the dates when those purchases were made, they will likely contain no information about weather conditions. The well-known relationship between weather and ice cream sales is a reflection of domain knowledge. To generate a good forecast, this knowledge must therefore be applied to the data for inclusion in the predictive modelling.

Simply put, domain knowledge is needed to decide what data can be used, how the data must be enriched and how the data should be transformed to support the modelling.

Domain knowledge is any additional information that may be available about a situation. For example, if there are gaps in the data, domain knowledge may explain that the sales process or production was halted for that period. Based on this insight, the data may now be treated accordingly, as the data is not really null or missing in the sense of having been omitted, but rather is null for a distinct reason. Domain knowledge also includes metadata. For example, when monitoring the sales of a product, one's main interest will be in the quantities sold and their selling price. However, metadata about staffing

levels at the point of sale can also provide information that helps with interpretation.

Most of the time it is also useful to apply domain knowledge to the way data preparation is done, for example, by making use of knowledge about the usual range and type of data items.

It is also important to rethink what is not documented in the data but should be there and might have an influence.[4]

This may include:

- changes to the marketing setup;
- changes to alert functions;
- old data that got lost during system updates;
- changes to the marketing strategy (eg more budget for lead generation);
- products that were changed at a certain point;
- raw materials from different suppliers that meet the same requirements but are not labelled as coming from a different source;
- the combination of staff members during different shifts (in production, at the point of sale, in the call centre);
- the fact that if a customer decides to do nothing (no click, no purchase, no complaint etc), no data is stored;
- the fact that if nothing unusual has happened, no data is stored; or
- the fact that the temperature in the production hall is colder or warmer than average.

These various kinds of information are likely to be stored somewhere, for example, in meeting notes, documentation or internal knowledge bases, but most of the time they are neither delivered with the data nor accessible in the metadata. For this reason, data scientists need to be in close contact with colleagues from relevant departments and those responsible for the data (data owners), to exchange and discuss surprising outcomes or results that lack an obvious analytical explanation.

The following example illustrates how first-hand knowledge can help an analyst. Consider a company that rents out motorhomes in Europe and needs to forecast the demand in the different countries. The regulations for motorhomes vary from one country to another. In some countries it is possible to park on public land for more than one night, while in others it is only allowed on camping sites or designated pitches. For motorhome travellers, this means that pitch fees must be added to the cost of the rental, and routes cannot be planned so freely.

Thus, the cultural context of the customer or the target region has an influence on the customer's decision.

A data scientist with this knowledge can therefore make targeted decisions in the model-building process that one without this knowledge would have missed.

Most of the time, good decisions need additive information that is not stored or recorded but is nonetheless present — perhaps in the memory of colleagues — and must be activated by asking questions. This can avoid wrong decisions and expensive mistakes, or at least missed opportunities.

ADDING AND REMOVING VARIABLES BASED ON DOMAIN KNOWLEDGE

Domain knowledge not only helps the data scientist to select the most relevant variables from a statistical point of view, but also makes it possible to add or remove variables by taking into account the implicit influences of business, culture, season and other factors.

Sometimes it makes sense to keep variables in the variable list, even if they are not necessary from a statistical point of view, because they are mostly (but not completely) explained by other variables. This is because the (small) unexplained part of these variables may add valuable insights to the model under domain knowledge aspects.

Variables can also be removed from the list when domain knowledge identifies spurious correlations — events that

coincidentally have a statistical relationship (positive or negative), but do not have an influence on each other (eg ice cream consumption and sunburn).

On other occasions, it may also help to create dummy variables based on domain knowledge and to introduce them into the model.

SUPPLEMENTING MISSING DATA WITH DOMAIN KNOWLEDGE

Because most data-mining projects use observational data collected from existing processes rather than well-designed experiments, missing data can be a common problem. To fix this, a distinction must be made between 'real missing data' and 'not stored information'. Domain knowledge helps make this distinction.

Typical examples for real missing data include missed birthday or age information in marketing problems or temperature or moisture measurements in technical datasets. Real missing data may be found in datasets generated from situations where the information itself most certainly exists in real life but for unknown reasons is not stored in the dataset. For example, every customer has an age and a date of birth, regardless of whether the dataset contains this information. The data may be missing as a result of the customer's explicit wish not to share the information, or because one process or another for obtaining the information did not work correctly. Missing information on temperature or moisture or such like, meanwhile, is very dependent on process errors or technical faults.

In the event of missing data being detected, imputation strategies may be used to replace the data with estimates. The estimation method differs depending on the business context and on other information available. Estimation may involve using mean or median figures (based on all data or just a relevant subset) or by a more complex method such as a regression equation or time series analysis. Cases with missing variables should only rarely be excluded.

An alternative way to replace real missing values is to use third-party knowledge. For example, a look-up table can be used to deduce gender from first names. For instance, Mary is very likely female and John almost certainly male. If such a table is not available, then one should be constructed from the existing dataset. When constructing such a table, it is essential to use the same coding as the existing dataset. For example, if the full dataset uses 1=male then the look-up table should also use 1=male. To combine the dataset containing missing values with the look-up table, the two tables should be merged using the first names as the key. It is also possible to create a look-up table from the salutation (eg Mr (Herr) or Ms (Frau), etc).

Information that has not been stored is very different from genuinely missing data. Rather than leave fields blank, it is better to enter a zero or some other value that represents the business knowledge behind it. These cases occur because most company databases only store things that have happened and not things that have not happened. For example, customers who have made a purchase create footprints in the database. This fact is stored in multiple tables, and one will be able to retrieve the date of purchase, the products purchased, the amounts purchased, the ticket price, the way the goods were ordered, paid, delivered, and so on. By contrast, for customers who have not made a purchase, the company database will say nothing. For some analyses, however, it is important to map the fact of 'not buying' in the dataset that will be used for data mining — especially if the majority of customers are not buying.

Instead of representing the 'not happened' case with a zero, one can also count the number of days since the last time something happened. Naturally, this kind of missing data must be replaced with a value that indicates the business meaning, but

beyond that, it should also fit the preferred method of analysis. For example, any 'no buy' may be represented by zero if the variable itself contains values (money) or amounts (pieces). If it contains 'days since', however, one cannot represent the missing data with zero because that would wrongly lead one to interpret that something has happened quite recently. Here it may be better to use the number of days since the last known activity (for example subscription for an e-mail newsletter) for estimation purposes or to substitute similar values that correspond to business rules. This may introduce unusually large values into the dataset, which may need to be transformed or considered as outliers.

HANDLING OUTLIERS WITH DOMAIN KNOWLEDGE

Outliers are another problem where domain knowledge provides support. Outliers are unusual values that show up as very different from other values in the dataset. They can be caused by process errors, or by unusual customer behaviour or other extraordinary occurrences, or by outstanding marketing events. In the case of process errors, it is possible to handle the outlier in a similar manner as one would a 'real missing' value; alternatively, if there are sufficient cases, one can reject the case that includes the outlier. Where it is possible to be certain that it is not a process error, then it is quite common to use techniques such as standardisation or binning or the quantile method to mitigate the effect of the outlier, or creating a dummy variable to indicate that the customer has reacted to the outstanding marketing event.

BINNING EMPOWERED BY DOMAIN KNOWLEDGE

To stabilise and improve predictive models, it is preferable to classify continuous variables, such as turnover, amount or days of purchase, into different levels. Using such a classification it is possible to emphasise more strongly the differences between levels, which is important from a business point of view.

This is particularly the case with variables that are conceptually non-linear. For example, it is important to know that a buyer belongs to the best 10 per cent of buyers, but the numerical distance between turnovers of say €2,516 and €5,035 is less important. Likewise, from a mathematical and statistical point of view, €1 is rather like €0, while from a business point of view, €1 indicates that the person has actually made a purchase, however small, and is therefore a better prospect than someone who has made no purchase at all.

This kind of binning technique can be used for all kinds of data. There are two ways to do the binning: under consideration of business rules and domain knowledge and under consideration of statistics and analytics. Binning techniques based on business rules and domain knowledge will usually have a bigger impact on model quality than using statistics alone.

Sometimes it is simply important to transfer stored information into meaningful information. For example, if the stored information is 'Berlin (Germany), 21.12.YYYY 18:00', may be translated to winter, dark, quite cold; if the stored information is 'Cape Town (South Africa), 21.12.YYYY, 18:00', may be translated to summer, light, warm. It is likely that these six new dummy variables — winter (0/1), summer (0/1), dark (0/1), light (0/1), cold (0/1) and warm (0/1) — will add more to the predicted model than using the basic date, time and location alone (see Figure 5).

The transfer of domain knowledge to the data using the described technique will create robust variables that are strong and meaningful, and that will lead to better models. The insights gained in this manner will drive innovative ideas on how to increase the business.

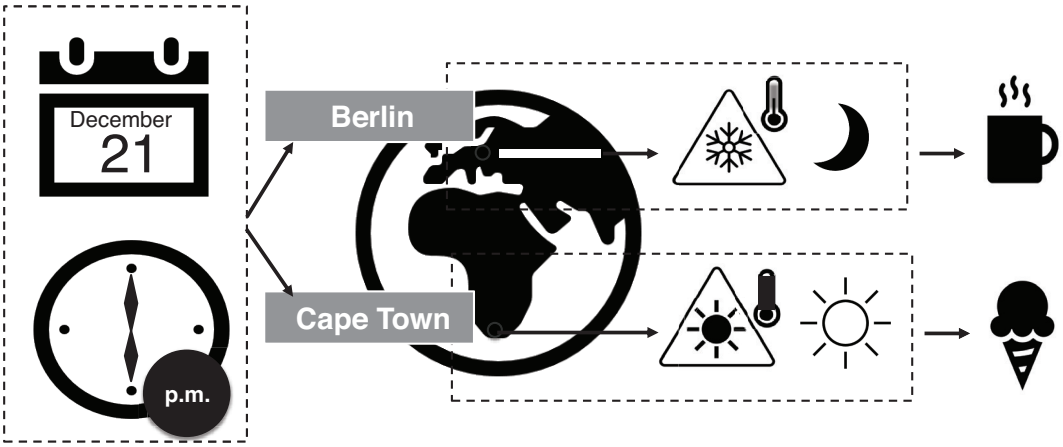

Figure 5: Dummy variables based on domain knowledge add value to predictive models

CONCLUSION

Data scientists should look to develop field-specific domain knowledge as well as knowledge regarding the more general aspects of their business. They need a clear view of what and how data has been recorded and archived,[2] and must understand that processes may have changed since the data were initially recorded.

Unwritten laws and historical events can have a big impact on products, product quality and the data stored. Analytics experts must maintain a strong relationship with the people preparing the data — when it comes to providing more insightful results, having well prepared data is more important than one's choice of analytical method.[2]

Data scientists also need to think further about what questions need answering, which areas need investigating and what insight can be gathered from the data. Often the original aims of the investigation could be expanded and developed to give improved insight and provide more valuable information. It is helpful to reflect on the task and to take into account the background and the business goal of those who are asking the questions, as they might not recognise when the task definition is vague or misleading for people with different backgrounds.[5]

Before moving forward, it is vital to invest time and effort to learn everything possible about where, and how, the analytical results should support the business or research goals.

Benefitting from the 4th industrial revolution requires both domain knowledge and technical skills.

References

1. Coleman, S. and Ahlemeyer-Stubbe, A. (2020) 'Monetizing industry 4.0 and IoT data for marketing and sales', in StatsRef Statistics Reference Online, Wiley, available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08269 (accessed 20th January, 2021).
2. Ahlemeyer-Stubbe, A. and Coleman, S. (2014) 'A Practical Guide to Data Mining for Business and Industries', John Wiley & Sons Ltd, Chichester.
3. Ahlemeyer-Stubbe, A. (2001) 'Analyseorientierte Informationssysteme = Datawarehouse', in Perner, P. (ed.) 'Data Mining, Data Warehouse, Knowledge Management: Proceedings of the Industrial Conference on Data Mining, Leipzig, July', pp. 12–28.
4. Scheideler, E. M. and Ahlemeyer-Stubbe, A. (2017) 'Quality control of additive manufacturing using statistical prediction methods', in Padoano, F., and Villmer, J. (eds) 'Production Engineering and Management', OWL University of Applied Sciences, Lemgo, pp. 3-12.
5. Ahlemeyer-Stubbe, A. and Coleman, S. (2018) 'Monetising Data — How to Uplift Your Business', John Wiley & Sons Ltd, Chichester.