John W. Foreman. Data Smart: Using Data Science to Transform Information into Insight. Indianapolis: John Wiley & Sons, 2014, 432 pages

Reviewed by Marcia Laugerman, Assistant Professor of Practice in Statistics, College of Business and Public Administration, Drake University

Subject Area: Business Analytics

This book provides an introduction into data science in a comfortable, casual, and at times humorous way. It speaks to various techniques used widely in the practice of data analytics in a conversational style. Each chapter provides examples that are interesting and easy to follow which makes the more complicated details of the techniques easier to comprehend. The author does a very good job of describing data analytics techniques in the solution of real world business problems. The book advances creative thinking about the use of analytics in business operations. In addition to years of analytics consulting with businesses such as Coca-Cola, Royal Caribbean, and Intercontinental Hotels, the author is also the Chief Data Scientist for Mailchimp, a leading email marketing company.

The introduction begins by defining data science as 'the transformation of data using mathematics and statistics into valuable insights, decisions, and products.' The book is useful for anyone wanting to understand existing techniques and get new ideas about how to apply them in data-driven decision management before hiring (sometimes expensive) consultants.

Instead of a book on statistical programming, the author demonstrates some very complicated procedures using Excel with an emphasis on Excel Solver. Staying away from a specific language is intentional but in the last chapter, the author gives R coding for most of the procedures. Some of the procedures require an advanced working knowledge of Excel, (just try working through the first chapter reviewing Excel techniques to see where you stand). Using Excel adds simplicity to the analytics where Excel 'excels'-like in visualizing data, summarizing data, mathematical, statistical, and financial functions, and pivot tables. However, some of the formulas linking multiple functions over multiple pages are quite tedious in Excel. Since most people are familiar with Excel, this book provides a more non-threatening starting place. The list of techniques covered is impressive and applicable, especially when one considers performing these in Excel. The first chapter promises a working knowledge in:

- Optimization using linear and integer programming
- Working with time series data, detecting rends and seasonal patterns, and forecasting with exponential smoothing
- Using Monte Carlo simulation in optimization and forecasting scenarios to quantify and address risk
- Artificial intelligence using the general linear model logistic line functions, ensemble methods, and naïve Bayes
- Measuring distances between customers using cosine similarity, creating k-nearest neighbor graphs, calculating modularity, and clustering customers
- Detecting outliers in a single dimension with Tukey fences or in multiple dimensions with local outlier factors

- Using R packages to “stand on the shoulders” of other analysts in conducting these tasks.

The author’s greatest strength is his ability to discuss techniques contemporaneously with memorable examples and humor. The Junior High school dance is a great way to understand clustering. Boys on one side, girls on the other, and chaperones at the top of the triangle. The clusters make no sense until the clusters are identified by gender and age. Calling Euclidean distance ‘as the crow flies’ is another example. Naïve Bayes is characterized by ‘the incredible lightness of being an idiot’ in its use for text classification. In describing ensemble models (finding creative ways to use training data repeatedly to create an entire ensemble of models), the author asks; “is it better to have a small amount of really good pizza or a lot of really bad pizza (p. 252)?” In this type of artificial intelligence implementation, the answer is often more models with less quality.

The book concludes with important considerations in the application of data analytics, emphasizing that often the difference between success and failure is not always the technique utilized but understanding how the problem is identified and defined. The analyst is encouraged to engage with the people whose challenges are being undertaking to learn the business’s processes and the data that is generated and saved in these business processes. By learning what processes handle existing problems and what metrics are used to gauge success, the analyst can make sure they are solving the right problem for the process. “Find ways to articulate analytics concepts within your particular business context. Push your management to involve analysts in planning and business development discussions (p 397).” In industry, analytics is a result-driven process where models are judged by their practical value, not their complexity. Often the best model is one that maintains a balance between functionality and maintainability to better utilize data in targeting, forecasting, pricing, decision-making, reporting, and compliance.