Texts in Computer Science

More information about this series at http://www.springer.com/series/3191

Richard Hill • Stuart Berry

Guide to Industrial Analytics

Solving Data Science Problems for Manufacturing and the Internet of Things

Springer

Richard Hill
Department of Computer Science
University of Huddersfield
Huddersfield, UK

Stuart Berry
Department of Computing
and Mathematics
University of Derby
Derby, UK

ISSN 1868-0941 ISSN 1868-095X (electronic)
Texts in Computer Science
ISBN 978-3-030-79103-2 ISBN 978-3-030-79104-9 (eBook)
https://doi.org/10.1007/978-3-030-79104-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Megan and Daniel.

Foreword

Industrialisation is an essential part of global economic development. Central to this development is technology; creating of new ways of doing things to enhance quality, repeatability and to discover new frontiers of value creation.

The automotive and aerospace industries have been prime movers in the advancement of the application of technology to improve lives and generate wealth. Both the outputs of these industries, and the environments in which they are produced, require a continual application and adaptation of technology to make our activities safer, more affordable and ultimately more sustainable, as the global community becomes aware of our collective need to reduce the consumption of increasingly scarce natural resources.

Thus, innovation is one of our primary tools to address current and future challenges. Innovation gives the ability to respond rapidly to emergent situations and to make reasoned sense of historical experience, so that we can learn from the past to inform the future.

Data has always been central to industry; without measurements, quantities, reporting and accounting, we would not have been able to make the advancements that have been witnessed through industrialisation.

However, it is more recent developments in computing technologies that are creating new ways to use data to create even more value and more advanced products.

Through widespread application of wireless sensor networks, embedded systems, cloud computing and ubiquitous high speed network infrastructure, we can identify hidden patterns in operational data, store and process vast quantities of data and constantly refine computational algorithms to search, categorise and predict new behaviours in a complex, inter-connected world.

This use of technology to collect, organise, process and consume data provides industry with the ability to monitor performance, automate decision making through condition monitoring and predictive maintenance, create seamless supply chain linkages through the close integration of industrial processes and logistics, leads us to discover and release value streams that were not seen prior to the adoption of analytics technologies. These technologies are key as we enter into a more model-based engineering (MBSE) approach to industrial innovation.

The 2017 UK Government Industrial Strategy[1] has been a catalyst of technology awareness. While the 'Tier 1' leaders of industry have understood that technology is central to their competitive advantage, other suppliers further down the supply chain have not always been able to keep pace with the early adopters.

Financial constraints such as limited investment have restricted the extent by which small and medium sized enterprises (SME) can explore the benefits of technological innovation until the technology itself becomes more affordable.

We are now at the beginning of an exciting era where technology is relatively inexpensive and the key differentiator between a business that is agile and fit for the future, and one that may struggle to remain sustainable, is the knowledge of how to use data—specifically the techniques of analytics—to maintain their competitive edge.

April 2021

Dr. Paul Needham, Ph.D. CEng FIET
Visiting Professor
University of Huddersfield
Huddersfield, UK

[1]https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future.

Preface

Overview and Goals

Technology is a key enabler of business, and as computation and storage costs become lower, what was once a vision of *computing as a utility* is now becoming the reality. Cloud computing models have illustrated how new business value and competitive advantage can be created from new ways of collaboration; inexpensive microprocessors and pervasive broadband networks are facilitating processing power that can be embedded into a constant stream of new applications.

As we start to see the possibilities of physical objects, that are inter-connected to share data, we can start to contemplate the potential of what an Internet of Things (IoT) environment might look like. From an industrial perspective, and especially manufacturing, there is the need to process and move physical objects to create business value.

As organisations strive to differentiate themselves from their competition, new ideas to increase sales revenue places hitherto unrecognisable demands upon the whole manufacturing supply chain.

What were once complex, but manageable challenges in planning, scheduling, production control and logistics, are rapidly becoming situations that are impossible to supervise without automation.

How we automate our industrial processes, to some extent assumes that we know what needs to be automated and that we have the necessary experience and skills to be able to bring the automation to fruition in a reliable way.

At the heart of any investigation into industrial operations is an understanding of:

1. what data is required;
2. what data is available;
3. what data processing needs to take place;
4. how to communicate the results of the analysis to a business audience.

Guide to Industrial Analytics: Making sense of data science for manufacturing and the Industrial Internet of Things is an attempt to address the need of organisations who can see the possibilities of an inter-connected industrial world but do not know how to make effective use of their data. It is commonplace for software vendors to sell 'black box' solutions that only solve one specific problem, yet many industrial challenges can be solved with some knowledge of specific techniques that are commonly utilised in the field of *data science*.

Often, guides to data science target audiences that are fluent in abstract mathematics. This serves little purpose for busy professionals who need to concentrate on their business needs.

This book has deliberately focused on the need to understand the practical application of data science techniques to solve industrial challenges, with minimal knowledge of mathematics required. Where the mathematics is essential, a detailed explanation is provided.

As such, the key objectives for this book include:

- to present an understanding of the fundamental approaches to analysing data that is commonly found in industrial environments;
- to understand the procedures and thinking around the selection and cleaning of industrial data;
- to demonstrate how we can apply different aspects of data science to discover interesting insight within data, using commonly available tools;
- to explore ways in which we can use existing data to make predictions about the future;
- to explore the ways in which visualisation can be used to enable the improved comprehension of industrial data;
- to understand the application of simple techniques to common situations, while also being aware of their limitations;
- to identify areas of further study in what is a fast-moving domain.

Target Audiences

The use of data to obtain new value and create opportunities for industrial businesses has a broad appeal. We have deliberately focused on delivering a book that shows how to apply data science techniques to industrial scenarios, and therefore the text is couched as a set of learn-by-doing exercises.

We have also taken a pragmatic stance in terms of the tools used to illustrate the examples. All software used is either freely available (open source) or is generally regarded as pervasive; it is likely that industrial organisations will have access to spreadsheets such as Microsoft Excel, or alternatives, for instance.

As such, *business leaders*, *industrial managers* and *supervisors* will find the combination of *just enough* mathematics and extensive practical explanations of value to them. Many traditional texts are long on theory and short on application. The know-how in this book will help them make more informed operational decisions, which in turn will improve the quality of data available for strategic planning.

Application developers who work on industrial enterprise IT systems will also be able to observe the type of analysis that industrial personnel want to do, and it is hoped that this book will inform the design and specification of updates and modifications to such systems infuture.

University instructors will find that this book is a suitably concise volume that can help get advanced undergraduate and postgraduate students applying data science techniques quickly. Many new university courses are including data science, or aspects of it as part of the curricula, but it is the real-world application of these techniques that is often lacking.

Finally, *technical consultants* and commercially oriented *researchers*, who work directly with industry to deliver tangible improvements, will find the collection of how-to articles for common scenarios of use to them in their business, especially the section on visualisation, in order to successfully communicate insight and conclusions to their clients.

Organisation and Suggested Use

This book is organised into three parts:

- Part I introduces the concepts of manufacturing analytics and data science.
- Part II describes a range of techniques and approaches to solving problems.
- Part III illustrates the application of methods and processes by way of industrial examples.

Guide to Industrial Analytics should be used as a comprehensive introduction to the use of data science techniques in real-world situations. Part 1 of the book provides the general foundation of the important concepts and is a good place to start for readers new to the topics.

Since our presentation of topics is rooted firmly in practicality, we recommend that Part I is reviewed by all readers. While there are countless texts on the specifics of data analytics, our presentation of the application of these techniques is relatively unique and there is much to be learned by reading about a topic or concept int he context of the industrial environment.

Part II should be seen as a walkthrough in the application of analytics techniques that have been proven to work. While many more exotic techniques exist, we have focused on approaches where there is the most return for the effort expended. These approaches have wide applicability and will significantly enhance the reader's ability to conduct useful and profitable industrial analytics.

Part III looks at the wider context of industrial acceptance of analytics, demonstrating that the barriers to successful adoption are not always limited by technical prowess.

A series of appendices present essential technical material to support the bulk of the text.

The book is designed to help readers acquaint themselves with practical techniques for dealing with industrial data, before becoming a reference text when the important processes are practised and understood.

For *university instructors*, we suggest the following program of study for a 12-week semester delivery pattern:

- Weeks 1–2: Part 1;
- Weeks 3–7: Part II;
- Weeks 8–11: Part III;
- Week 12: Assessment.

Part I explores the context of industrial data, how it is used and what we can achieve with it. It provides a practical definition of data science and explores how we might apply techniques to different situations. There is also an introduction to the tools that can be used to perform the analysis.

Part II is a more in-depth look at a range of techniques that we use to find insight from data. Using a tutorial-based approach, there are specific examples that demonstrate data science skills in practice. There are also exercises for the readers to complete and reinforce their own knowledge.

Part III demonstrates the application of approaches and techniques to real life. These examples help illustrate how we can combine different techniques to solve a particular industrial query.

Learning Activities

Each chapter concludes with a set of review questions and learning activities that make specific reference to the material in each chapter. There is also an additional set of more open questions that will require further investigation from the reader. Such questions help embed the material that has been learned, so that it can be applied to a number of different situations. These questions will be useful to university instructors who can set them as homework activities outside of class.

Hands-on Exercises

Much of the application of this work requires proficiency in the use of tools. Much of the *mystique* of data analytics is knowing what tool works in what set of circumstances and we have deliberately focused upon the use of tools that are commonly available in the industrial environment. Exercises are used throughout the book to illustrate not only the 'what' but also the 'how' of data analytics. Our intention is for readers to develop sufficient skills to use the techniques as tools when they are faced with an industrial analytics job in the future, hence the strong thread of learning-by-doing.

Acknowledgements

The authors would like to express their gratitude to the many organisations and business owners from the Midlands and Yorkshire regions of the UK, for their cooperation and engagement with research and consulting projects over the past 25 years. Much of the insight within this book has come from practical experience, experimentation, evaluation and the general fieldwork of being involved in industrial operations.

In addition, the chapters explaining data analysis and statistics have been developed over 12 years of teaching undergraduate students, with particular contributions from Richard Self and Dr. Pritesh Mistry.

Huddersfield, UK — Richard Hill
Derby, UK — Stuart Berry
April 2021

Contents

Part I Introductory Concepts

1 An Introduction to Industrial Analytics 3
1.1 What Is Analytics? 3
1.2 Breaking Boundaries 4
1.2.1 The Industrial Internet of Things 6
1.2.2 Disruption Means Change 8
1.3 Industry 4.0 9
1.4 Opportunities for Smart Businesses 11
1.5 What Is Data Science? 12
1.6 Why Do We Need Data Science? 13
1.7 A Process for Data Science 14
1.7.1 Data Preparation 15
1.7.2 Data Exploration 17
1.7.3 Model Selection 17
1.7.4 Evaluation 18
1.8 Do We Need Machine Learning for Industrial Analytics? 19
1.9 Learning Activities 19
References 20

2 Data, Analysis and Statistics 21
2.1 Introduction 21
2.2 The Need for Analysis and Statistics 21
2.3 Qualitative and Quantitative Data 22
2.4 Data Terminology 23
2.5 Data Quality 23
2.6 Scales of Measurement 25
2.6.1 Nominal Data 25
2.6.2 Ordinal Data 28
2.6.3 Interval Data 28
2.6.4 Ratio Data 28

2.7 Central Tendency . . . 29
2.7.1 Mean . . . 29
2.7.2 Median . . . 30
2.7.3 Mode . . . 30
2.8 Dispersion . . . 31
2.8.1 Range . . . 32
2.8.2 Interquartile Range . . . 32
2.8.3 Variance . . . 32
2.8.4 Standard Deviation . . . 33
2.8.5 Frequency . . . 33
2.9 Histogram . . . 35
2.9.1 Cumulative Frequency Graph . . . 36
2.10 Shape of the Data . . . 36
2.10.1 Normal Distribution . . . 37
2.10.2 Uniform Distribution . . . 42
2.10.3 Bimodal Distribution . . . 42
2.10.4 Skewed Distributions . . . 43
2.11 Visualising Data . . . 44
2.11.1 Pie Charts . . . 45
2.11.2 Bar Charts . . . 47
2.11.3 Line Charts . . . 49
2.11.4 Scatter Plots . . . 50
2.12 Learning Activities . . . 50
References . . . 51

3 Measuring Operations . . . 53
3.1 Introduction . . . 53
3.2 Using Assumptions . . . 54
3.3 Operations Concepts . . . 55
3.3.1 Cycle Time . . . 55
3.3.2 Lead Time . . . 57
3.3.3 Takt Time . . . 57
3.4 Using Concepts to Understand Systems . . . 57
3.5 Resource Utilisation . . . 59
3.6 Learning Activities . . . 62
References . . . 63

4 Data for Production Planning and Control . . . 65
4.1 Historical Attitudes Towards the Use of Data . . . 65
4.2 Need for Data Within the Production Area . . . 66
4.3 Planning Problems Resulting from Lack of Appropriate Data . . . 67
4.4 Planning with Appropriate Data . . . 68
4.5 Need for Optimality in Production Control and Scheduling . . . 69

4.6 Deriving Generic Models for Planning and Control 75
4.7 Production Planning in Manufacturing: Small Case Study Results 78
4.8 Planning and Control in the Case Study Firms 79
4.9 Manufacturing Production Systems in Case Study Firms 80
4.10 Summary 81
4.11 Learning Activities 82
References 83

Part II Methods

5 Simulating Industrial Processes 87
5.1 Understanding Business Operations 87
5.2 Queues and Queueing 88
5.3 Modelling an Industrial Process 91
5.4 Designing a Process Simulation 92
5.5 Building the Simulation in Ciw 93
5.6 Confidence 96
5.7 Conclusion 98
5.8 Learning Activities 99

6 From Process to System Simulation 101
6.1 Simulating Industrial Systems 101
6.2 Example: Joinery Manufacturer 102
6.3 Building the Simulation 103
6.4 Managing Resource Utilisation 111
6.5 Product Mixes 113
6.5.1 Sash Windows 114
6.6 Conclusion 124
6.7 Learning Activities 124

7 Constructing Machine Learning Models for Prediction 127
7.1 Introduction 127
7.2 Data and Prediction 128
7.2.1 Example 1: Job Time Prediction Under Varying Demand 129
7.3 Assessing the Predictive Power of a Model 133
7.3.1 Root Mean Squared Error (RMSE) 133
7.3.2 Mean Absolute Error (MAE) 134
7.3.3 Mean Absolute Percentage Error (MAPE) 134
7.3.4 Coefficient of Determination (R^2) 134
7.3.5 Underfitting and Overfitting 134

7.3.6 Cross-Validation 135
7.3.7 Learning Curves 136
7.3.8 Validation Curves 136
7.4 How to Improve Model Accuracy 137
7.4.1 Feature Selection 137
7.4.2 Example 2: Improving the Model with Additional Information (Multiple Regression) 138
7.4.3 More Data 139
7.4.4 Compare Models 140
7.4.5 Example 3: Multiple Job Types (Model Comparison) 140
7.5 Generating Data Via Simulations 141
7.5.1 Example 4: Simulating Data Under Uncertainty 142
7.5.2 Kernel Density Estimation and Sampling 145
7.6 Worked Examples in R 145
7.6.1 Linear Regression Model 145
7.6.2 Multiple Regression Model 148
7.6.3 Cross-Validation 149
7.6.4 KDE Estimation of Distribution 150
References 151

Part III Application

8 Case Study: Confectionery Production 155
8.1 Introduction 155
8.1.1 Company Organisation 156
8.1.2 Production 156
8.2 Hard Boiled Confectionery Organisation and Planning 158
8.2.1 Unit Operation 159
8.2.2 Scheduling 160
8.2.3 Model to Determine Optimal Long Term, Monthly Production Plans 160
8.2.4 Implementation of Monthly Planning 164
8.2.5 Allocating Production Pairs 166
8.2.6 Implementation of Daily Pair Selection 169
8.3 Impact of Lack of Information on Company Profits 172
8.4 Conclusion 173
8.5 Learning Activities 174
8.5.1 Machinery and Staffing 176
8.5.2 Sales Data 177

9 Minimum Information Set for Effective Control 183
9.1 Information Flows Within an Organisation 183
9.2 Deriving Minimum Information Requirements 186

9.3 Order Book-Based Systems . . . 186
9.3.1 Deriving an Order Book (OB) Forecasting Model Where Orders for New Jobs Arrive Randomly in Time . . . 189
9.3.2 Variability in the Final Product Introduced at All Stages . . . 190
9.3.3 Simple Order Book System . . . 191
9.3.4 Enhanced Order Book Models . . . 192
9.3.5 Forecast Accuracy/Validation . . . 194
9.3.6 Extending the Investigation by Including Data from All Stages . . . 195
9.3.7 Variability Added (only) at the Final Stage . . . 196
9.3.8 Conclusion: Order Book (OB)-Based Approaches to Forecasting . . . 198
9.4 Work Book (WB) Systems . . . 199
9.5 Evaluating WB and OB When Stage Productions Have Been Balanced . . . 204
9.5.1 Conclusions and Recommendations MIR . . . 205
9.6 Data Requirements for Planning and Control . . . 208
9.7 Minimal Information in Flow Shops with CONWIP Control . . . 208
9.7.1 Flow Shops with More Than One Like Processor at Each Stage . . . 210
9.7.2 Job Shops . . . 211
9.7.3 Effect of Growth on Planning and Control in a Flow Shop . . . 212
9.7.4 Results from Three-Stage Models . . . 212
9.7.5 Results from 10-Stage Models . . . 213
9.8 Conclusion . . . 214
9.9 Learning Activities . . . 215

10 Business Adoption of Analytics . . . 217
10.1 Introduction . . . 217
10.1.1 Intelligent Manufacturing . . . 217
10.1.2 Compounded Challenges for SMEs . . . 218
10.1.3 Regional Challenge . . . 218
10.2 A Model of Engagement . . . 219
10.2.1 Proving the Return on Investment . . . 220
10.2.2 Digital Enablers Network (DEN) . . . 220
10.3 University Capability . . . 222
10.3.1 Case Study: DEN in Action . . . 222

10.4 Discussion 223
10.4.1 Benefits to SMEs 224
10.4.2 Benefits to den Members 224
10.4.3 Benefits to Academia 224
10.4.4 Human Factors 225
10.5 Conclusions 225
10.6 Future Work 226
References 227

Appendix A: Statistics 229

Appendix B: Simulation Library—Ciw 243

Appendix C: Production Planning Programs in MS Excel 245

Appendix D: Scheduling Jobs Through a Workshop 263

Index 273

Contributors

Stuart Berry University of Derby, Derby, UK

James Devitt University of Huddersfield, Huddersfield, UK

Richard Hill University of Huddersfield, Huddersfield, UK

Sam O'Neill University of Derby, Derby, UK

Abbreviations

3DM	Data-driven decision-making
ACID	Atomicity, consistency, isolation and durability
AD	Automatic differentiation
AE	Autoencoder
ANN	Artificial neural network
AOSD	Aspect-oriented software development
API	Application programming interface
AQL	Annotation query language
ARD	Automatic relevance determination
ASR	Automatic speech recognition
BDA	Big data analytics
BNN	Binary neural net
BPTS	Back-propagation through structure
BPTT	Back-propagation through time
CART	Classification and regression trees
CCA	Canonical correlational analysis
CEP	Complex event processing
CNN	Convolutional neural network
COCO	Common objects in context
COTS	Commodity off-the-shelf
CPPN	Compositional pattern-producing network
CQL	Cassandra query language, cyber-query language, common/contextual query language
CRF	Conditional random field
CTC	Connectionist temporal classification
CV	Cross-validation
DAD	Discover, access, distil
DAG	Directed acyclic graph
DBN	Deep belief network
DCGAN	Deep convolutional generative adversarial networks
DHSL	Distributed Hadoop storage layer
DNN	Deep neural network
DT	Decision tree
EBM	Energy-based model

ECL	Enterprise control language
EDA	Exploratory data analysis, event-driven architecture
ELU	Exponential linear unit
EPN	Event processing nodes
ERF	Extremely random forest
ESP	Enforced subpopulations
FUSE	Filesystem in userspace
GA	Genetic algorithm
GAN	Generative adversarial network
GBM	Gradient boosting machine
GEOFF	Graph serialisation format
GMM	Gaussian mixture model
GRU	Gated recurrent unit
HAM	Hierarchical attentive memory
HAR	Hadoop archive
HMM	Hidden Markov model
HPCC	High performance computing cluster
HPIL	Hadoop physical infrastructure layer
HTM	Hierarchical temporal memory
IDA	Initial data analysis
IIoT	Industrial Internet of Things
IoT	Internet of Things
JAQL	JSON query language
JSON	JavaScript object notation
KFS	Kosmos file system
KNN	k-nearest neighbours
KPCA	Kernel principal component analysis
KSVM	Kernel support vector machine
LOOCV	Leave one out cross-validation
LReLU	Leaky ReLU
LSTM	Long short-term memory
LTU	Linear threshold unit
LZO	Lempel-Ziv-Oberhumer
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MCMC	Markov chain Monte Carlo
MDM	Master data management
MDP	Markov decision processes
ML	Machine learning
MLP	Multi-layer perceptrons
MSA	Microservices architecture
NB	Naive Bayes
NEAT	Neuro-evolution of augmenting topologies
NLP	Natural language processing
NN	Neural network

NTM	Neural turing machine
OCR	Optical character recognition
OLAP	Online analytical processing
OLS	Ordinary least squares regression
OLTP	Online transactional processing
PAC-MDP	Probably approximately correct in Markov decision processes
PCA	Principal component analysis
PMML	Predictive Model Markup Language
PReLU	Parameterised ReLU
RBM	Restricted Boltzmann machines
RDD	Resilient distributed database
ReLU	Rectified linear unit
ResNet	Residual neural network
RF	Random forest
RL	Reinforcement learning
RMSE	Root mean squared error
RNN	Recurrent neural network
RNTN	Recursive neural tensor network
RTRL	Real-time recurrent learning
RVM	Relevance vector machine
S4	Simple scalable streaming system
SANE	Symbiotic adaptive neuro-evolution
SD	Standard deviation
SGD	Stochastic gradient descent
SIFT	Scale-invariant feature transform
SOA	Service-oriented architecture
SRN	Simple recurrent network
SVD	Singular value decomposition
SVM	Support vector machine
TDA	Topological data analysis
TF	TensorFlow
TFIDF	Term frequency inverse document frequency
UDAF	User-defined aggregate function
UDTF	User-defined table-generating function
UIMA	Unstructured information management architecture
VC	Vapnik Chervonekis dimension
VLAD	Vector of locally aggregated descriptors
W3C	World Wide Web Consortium
WFST	Weighted finite-state transducers
XML	Extensible Markup Language
YARN	Yet another resource manager