Big Data and Business Analytics Collection
Mark Ferguson, *Editor*

Introduction to Business Analytics

Majid Nabavi
David L. Olson

BEP BUSINESS EXPERT PRESS

Introduction to Business Analytics

Introduction to Business Analytics

Majid Nabavi and David L. Olson

BUSINESS EXPERT PRESS

Introduction to Business Analytics

First published in 2019 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-94944-327-1 (paperback)
ISBN-13: 978-1-94944-328-8 (e-book)

Business Expert Press Big Data and Business Analytics Collection

Collection ISSN: 2333-6749 (print)
Collection ISSN: 2333-6757 (electronic)

Cover and interior design by Exeter Premedia Services Private Ltd., Chennai, India

First edition: 2019

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

Abstract

This book is intended to present key concepts related to quantitative analysis in business. It is targeted to business students (both undergraduate and graduate) taking an introductory core course. Business analytics has grown to be a key topic in business curricula, and there is a need for stronger quantitative skills and understanding of fundamental concepts.

Keywords

business analytics; forecasting; hypothesis testing; knowledge management; regression; statistical sampling

Contents

CHAPTER 1

Business Analytics

Business analytics refers to the use of quantitative analysis to support managerial decision making. It is concerned with the process of managerial decision making, as well as the tools used to support it (management science). Information is a valuable asset requiring careful management, to make sure that businesses are able to accomplish their missions, no matter what those might be. In the agricultural domain, the production, processing, and marketing of food to keep us all going involves many participants, although there are a relatively small number of major producers such as ConAgra, ADM, Kellogg's, and so on. These major producers have massive data processing challenges, concerning weather, markets, production of many commodities across the world, and many types of demand. In the retail domain, Wal-Mart and its competitors deal with many products coming from many places, and stock these products in many retail outlets, served by a massive distribution network. Keeping on top of data is crucial to Wal-Mart's success, served by what is probably the world's largest data processing system. Distribution requires a massive transportation network. One element is the airline industry, which has to keep track of many customers seeking to travel from one location to another. Governments also have massive data needs. Providing protection to citizens requires collecting and processing data on airline traffic, shipping, road networks, population movements, economic activity, and risks of nature in the form of weather, volcanic activity, tsunamis, and so on.

Information can be viewed in a variety of ways. You might think of raw data as an initial form that humans need to process into understanding through various stages. But classifying these stages is probably unnecessary—our world today is flooded with data from a variety of forms. For almost centuries, the U.S. government has collected census data (as the

Romans did millennia ago). In the last century, a sophisticated system of economic data collection was developed in the United States, followed by OECD in Europe, and UN data of various types. Businesses also collect data, seeking to gain competitive advantage by understanding what the opportunities are in their markets. Science has been in favor of open data, shared so that humanity as a whole can gain understanding more rapidly. The medical scientific field consists of a complex network of researchers seeking cures to cancer and other evils. The business domain overlaps this scientific endeavor in the form of the pharmaceutical industry, resulting in an interesting dichotomy of interest between open sharing of information for the sake of progress versus intellectual property protection to further pharmaceutical profitability.

There are two major developments in recent years that have revolutionized the field of information. Supply chain networks have evolved, enabling linking a multitude of vendors from practically anywhere across the globe with businesses adept and processing or manufacturing products. Their production expertise can often be applied in many different locations, enabling them to find cheaper labor to make products, transported over a sophisticated distribution network. It is astounding to consider how fruits from Central or South America, grains from Africa as well as the Midwestern U.S., and meats from Australia and Argentina can find their way to our groceries in Europe or the United States. The resulting supply chain networks usually find it worthwhile to be open in sharing data with their supply chain partners, much as keiretsus and chaebols have operated in eastern Asia.

A second major development is the explosion of social media, which has led to linkages of people around the globe. Many countries find far more people using cell phones than landlines, and even places like the United States where landlines have been in place for over 100 years now see most people abandoning the old systems for cheaper and more mobile cell phones. These devices, along with a plethora of other platforms such as smartphones and tablets supported by free access to the World Wide Web, enable people to talk to each other over Facebook and to purchase products from organizations such as Amazon. This has led to masses of data being accessible to businesses to collect and analyze (a business form of Big Data).

Business Decision Making

Science seeks to gain complete understanding of whatever topic is under consideration, identifying the entire system of interacting parts, how they relate to each other and how this leads to outcomes. The field of physics has successfully been able to send rockets (some occupied by humans) to the moon, and more impressively, to return safely. Geology has studied the components of the earth and scientists are able to identify properties useful to growing crops in particular locations, as well as identify likely places for oil discovery or the discovery or rare earth elements. Chemistry has been able to develop combinations of atoms to propel vehicles, to blow things up, and to deliver pills making us feel better. Biology seeks to understand how organisms go through their life cycles.

Science has accomplished a great deal, but not that humans have not gained perfect understanding of anything. The degree of how well bodies of scientific theory can explain natural phenomena varies. In physics and geology, the scientific process has accomplished a great deal. Chemistry has also seen many useful gains, although the interactions of various drugs on humans are far from understood. Biology includes even greater uncertainty, and medical science has a long way to go to master human and other animal systems. When you get to economics and other human behavior, the mysteries deepen.

Business has always involved the interactions of many people. These people make assumptions, to cope with the complexities of their lives. Making assumptions is a necessary part of theory building necessary to surviving everyday life. But many times assumptions are based on false speculation, resulting in incorrect understanding and poor decisions. The contention we propose is that gathering data and measuring what is going on in businesses can lead to better understanding, and consequently, better decision making. This is not expected by any means to be a cure-all. No matter how much we think we know, there will be mistakes in our theories of cause and effect, as well as our understanding what is currently going on. But we have to keep trying.

LaPlace once contended that if he knew the starting conditions, he could calculate how anything would behave. This represents a type of reductionism, which pure scientists often adopt. It is a Newtonian view of strict causality, which implies that if you can't measure and explain, it isn't scientifically understood. This view complies with the idea of determinism, that everything that unfolds in the world is describable by natural law.

This idea is not necessarily wrong, but in the domain of business seems impractical, either due to the complexity involved in our world, or the continual change endemic to natural as well as human activity. Thus it is necessary to view life more flexibly, to understand that systems are complex and change, and that what has happened in a particular context in the past is not necessarily going to be the same in the future. Business has to face changes in laws and regulations, in societal attitudes regarding what is acceptable behavior, and the high levels of uncertainty involved in markets.

Thus there is not "only one best way" in any business decision context. Management is getting things done through people, and whenever you have people involved, you find changes in mood in individuals, and different attitudes across people, and different market behavior for different profiles of customers. Management thus has to combine a scientific-like process of theory verified by observation along with an art of getting along with a wide variety of people. Managerial decision making is the most fundamental function of management, and a very difficult task where demanding stakeholders are never satisfied.

Scientific Method

Science is a process, seeking to carefully study a problem and evolve over time with a complete mathematical description of a system. This seeks to be completely objective, measuring everything accurately and developing theories with no emotional bias. This scientific approach, as we have alluded to before, has served humanity well in many environments. We think it seems to do better when human choice is not involved, as we humans have always reserved the right to change our minds and behavior.

The scientific method can be viewed as a process consisting of the following elements:

- Define the problem (system)
- Collect data

- Develop hypotheses of cause and effect in the interaction of system elements
- Test these hypotheses
- Analyze results
- Refine our mental models or theories of cause and effect

This scientific method has led to impressive understanding of the laws of physics. Astronomers continue to expand our understanding of the universe, through more powerful telescopic tools as well as spectral analysis of elements. Einstein was able to formulate a mathematical expression of the relationship between energy, mass and light. USSR cosmonauts and NASA in the United States were able to send rockets into orbit, and to enable men to walk on the moon.

It does make sense to try to apply this same mindset to business. Businesses are complex collections of individuals with different backgrounds, training, and roles, each with their own set of ambitions and agenda. Thus there will often be differing theories, biased by perceptions of self-interest, among the people in any organization. We expect that those that operate more objectively will prevail (although life doesn't guarantee that), and that businesses will do better if they are guided by leaders seeking to be as scientific as they can be, given that they understand that economic activities are very complex and changeable.

Management Decision Process

Keeping in mind that we don't expect that you can be completely scientific in business, a sound objective approach would be expected to serve businesses better than a chaotic decision making environment that operates at random. Thus an analogy to the scientific method outlined earlier might be:

- Define the decision problem (what products to carry)
- Search for data and information (what demand has been observed in this area)
- Generate alternative actions (gather prospective vendor products and prices)

- Analyze feasible alternatives (consider price, quality, and delivery)
- Select best action (select products and vendors)
- Implement (stock your outlets)

This process provides a means to be as scientific as possible within the confines of the chaos and uncertainty of business. Defining decision problems is often imposed by the job business people have. Getting data is critical. Organizational information systems are designed to provide reports that were expected to provide key information enabling everybody in the organization to do their job. But these reports rarely include everything that matters, so employees have to supplement ERP reports with other information. Observation and talking to people is a good source. But we can also go find other information, on the Web, in libraries, even creating experiments to try to understand the systems we have to deal with. Statistical analysis provides valuable tools to monitor performance.

Generating alternative actions is often a matter of creativity. Part of the job of management is to monitor your responsibilities to identify where action is required. This is hard, because it is human nature to see maybe nine problems for every real problem. Systems often self-correct, and micromanaging can be highly detrimental. Experience is extremely valuable in this context, learning when to take action and when to leave things alone. If action is taken, there are many things that could be done. We will look at models as means to anticipate expected outcomes, or in special circumstances, suggest optimal solutions. They are valuable tools for analysis of alternative actions.

Selecting the action to take should not be surrendered to models. Every model involves assumptions, and the models that appear the most powerful usually involve the most assumptions. Thus human judgment needs to be the final decision maker. Models provide useful input, and means to experiment. But we still live in a world where humans are in charge of computers.

Once action is taken, the results need to be monitored. If you make a decision to raise prices on your product, you need to monitor the impact in demand (which probably will go down, but you won't know by how much until you observe and measure).

Management Science

Management science is the field of study seeking to apply quantitative models to common management decisions. It is a parallel field to operations research, which developed in the engineering field. The modeling tools used are common to both fields. These models allow decision makers to experiment, intending to learn more about their operations.

Management science involves purpose-oriented decision making, following the management decision process we have just outlined. Systems have to be understood, including relationships between various resources available to managers, which they apply to accomplish organizational goals. Management science is a toolbox of quantitative methods available. Decision making units need to understand problems, monitor data, and then select one or more management science methods to try to better understand what to expect.

Table 1.1 outlines one view of the primary tools available from management science.

Table 1.1 Management science techniques

Category	Technique	Example use
Statistics	Measure	Monitor performance
Probabilistic	Queuing	Waiting line performance
	Inventory analysis	EOQ, ROP
	Markov chains	Marketing
Simulation	Monte Carlo	Distributions, system performance
	System simulation	Complex waiting line performance
Optimization	Linear programming	Best solution
	Transportation	Moving things the cheapest way
	Assignment	Assigning tasks the best way
	Nonlinear	Economic models
	Network models	Special cases
	Critical path	Project scheduling
	Dynamic programming	Contingent behavior
	Nonlinear programming	Chance constrained models
Decision theory	Decision trees	Probability
Game theory	Consider actions of others	Competitive environments

Statistics can be viewed as the science of collecting, organizing, analyzing, interpreting, and presenting data. It is a fundamental quantitative set of techniques used to measure things. Almost every input to the other models in Table 1.1 are obtained through some level of statistical analysis. Even when human opinion is used to generate input, the impact should be measured statistically.

There are a number of probabilistic modeling techniques. In production systems, or in service business, waiting lines are commonly encountered. Inventory modeling is a fundamental probabilistic component of operations management, with demand being notoriously probabilistic. Analytical models exist for given sets of assumptions, just as they are in queuing models, but simulation is attractive in that it allows any number of assumptions to be included in models. In production systems, arrival rates may be fairly constant, as in assembly lines. But in service systems, such as bank teller windows, arrival of business tends to be highly variable. The expected performance in terms of waiting time and waiting line length can be predicted through queuing models. Markov chains are another form of probabilistic model, where various states can be occupied by entities, transferring to other states at various rates. Overall equilibrium performance can be predicted through these models.

Simulation is a very useful general field that can be used to model just about any system. This enables realistic models with probability distributions allowable for system entities. Unfortunately, the more reality that is modeled through probabilistic elements, the more difficult it becomes to make sense of the output. Probabilistic models such as queuing work when only a few rigid assumptions are made. Simulation allows the modeler to include as many assumptions as desired.

Optimization is the most powerful type of management science model. The standard is linear programming, where any one objective function can be optimized subject to a set of linear constraints. This has proven highly valuable in many contexts, to include petroleum refinery planning, or factory product mix decision making. The downside is that linear programming generates optimal solutions only if assumptions relative to linear functions and parameter certainty are met. If these assumptions are not appropriate, the resulting linear programming solution can be radically inferior. Linear programming also assumes that continuous variable values are allowed, although refinements in the form of integer or zero-one programming are available if that assumption is not appropriate.

There are many modifications to the basic idea of optimization modeling. Transportation models are a special case of linear programming, where there are sources and destinations, with different costs over each connection. This could be modeled as a linear program, but a transportation algorithm is much more efficient, allowing modeling much larger problems. Assignment models are a special case of transportation algorithms, where the sources and demands are all one.

Network models can be modeled for specific problems such as identifying the shortest path to visit each of a set of locations, or finding a lowest-cost way to connect a network of elements, or maximum flow models seeking to find the highest volume of flow over a network. Critical path models are useful in project management, identifying the earliest finish time schedule for a series of project activities with given durations and predecessor relationships.

Dynamic programming optimizes, but in a sequential form of modeling, enabling dealing with problems that would otherwise involve too many combinations. Nonlinear programming is just like basic linear programming, with nonlinear functions allowed. This is useful in some contexts such as chance constrained modeling, where the assumption of certainty is replaced with statistical distributions. Nonlinear models are harder to optimally solve, but software has become widely available to support it.

Decision theory includes two useful categories of models. Decision trees provide a means to incorporate combinations of decision maker choices with probabilistic outcomes. This broad category of model could also include multiple attribute decision making, to include utility theory, analytic hierarchy process, and other techniques to make selection choices under conditions of trade-offs. Game theory is a class of modeling dealing with competitive situations, considering the actions of others.

Knowledge Management

Knowledge management (KM) is a broad field including identification, acquisition, storage, and retrieval of information to aid decision making.

These four functions are processes of KM. Identification requires focusing on the information important in supporting organizational decision making, selecting the best available measures to support these decisions. Knowledge acquisition involves getting the data providing these measures, which can involve aggregating data from internal systems such as ERP, extracting data from governmental or commercial sources for data external to the organization, and even conducting research to obtain more specific data. Storage is usually an information systems or information technology task, supplemented by individual databases. And what is entered into storage needs to be retrievable.

KM is process oriented, thinking in terms of how knowledge can be acquired, as well as tools to aid decision making. Rothberg and Erickson's (2005) framework defined data as observation, which when put into context becomes information, which in turn can be processed by human understanding to become knowledge. Big data can be a source used to generate insights, innovation, and business value by providing real-time measures of performance, more timely analyses based on more complete data, and can lead to sounder decisions.

Big Data

KM today is characterized by the existence of big data. The advent of i-phones and other hand-held devices has led to an explosion in data. Our culture has become obsessed with sharing many details about ourselves. Some of this self-centered desire to share everything about ourselves with the world has been found useful to many retail organizations. There also is a great deal of useful information generated from e-business activity. Thus we have big data, too massive to store on a single server, too unstructured to fit within standard spreadsheet formats, continuously generated, with little structure (Davenport 2014). This big data explosion has had highly important impact on KM, offering many opportunities to business organizations and to identity thieves.

Three aspects of big data found important are volume, velocity, and variety:

- Volume is an important aspect of KM, as streams of data arrive in real time from cash registers. Large organizations such as Wal-Mart have found it worthwhile to capture this information, aggregate it, providing capability to generate customer profiles to enable real-time marketing opportunities custom tailored to milk the maximum revenue stream from each source. This information can also be used to manage inventories, and to deal with vendors.

- Velocity is important to enable real-time response. One of the most voluminous types of data is the weather data generated by satellites, streamed back to earth-bound computers, which need to process this information and feed useful information to weather reporters throughout the world. Military operations also have high velocity information that needs to be made sense of to enable rapid decisions concerning targeting and other military applications. Retail business also needs to be able to operate in real-time, which requires high-velocity capabilities.
- Variety is important in many applications. Social media generate data useful to retail businesses. This social media data consists of many data formats, to include networks of links, photographic data, movie data, and so on. The medical industry has become a major part of the global economy, with even more complex data formats, to include MRI data, DNA data, and ever-evolving new format types.

The Computer System Perspective

The role of information systems is to provide organizational leadership with an accurate picture of what is going on in their areas of responsibility. Management information systems have evolved to sophisticated enterprise systems, integrating diverse reporting elements to provide comprehensive and systematic reporting. But there has always been an evolutionary aspect to organizational information, with individual elements having their own needs. Within the field of enterprise systems, there are giant firms such as SAP, Oracle, and Infor that provide top-scale support. These are supplemented by a level of commercial systems specializing in areas such as transportation management systems, warehouse management systems, manufacturing execution systems, advanced planning systems, and so on. There will always be new software tools developed and marketed that offer more focused and affordable service to their clients. This includes focused systems that were once referred to as decision support, providing dedicated systems to provide individual decision making entities with access to data, modeling, and reporting tools for specific decision problems.

Among the many computer systems supporting decision making are tools such as Tableau, which provides a useful database by county or zip-code, with demographic data, that can be used to visually or tabular display demand for specific products. These software tools emphasized the visualization aspects of information analysis, providing highly useful marketing support.

The Analytics Perspective

The analytics perspective involves various forms of management science, which has been around since World War II, and in the field of inventory management, longer than that. This includes the need to consider sustainability in this era of global warming, as well as risk management. Davenport (2013) reviewed three eras of analytics. In the first, business intelligence, focused on computer systems such as decision support systems harnessing custom-selected data and models. In the early 21st century, big data generated from Internet and social media provided a second focus. Davenport saw a third era involving a data-enriched environment with online real-time analysis.

The Internet of things provides an additional source of big data. Just as people can communicate through text, e-mail, and other forms of communication, machines communicate with each other. In the health care industry, Fitbits and other personal monitoring devices generate data that could conceivably link to personal physicians. The problem physicians would have coping with this potential flood of data is thought provoking. How people ever survived until 2010 is truly a wonder. Out of massive quantities of data, only a miniscule bit is germane. Of course, signals are sent only when critical limits are reached, but if the system scales up to include the majority of the billions of people inhabiting the earth, it would seem that some means of management of data volume would grow in importance.

Modeling Decision Environments

The point of analytical modeling is to construct an abstraction of a real system capturing the key elements involved, and then solving the resulting model to arrive at a way to improve this real system. Models can take many forms, to include maps, organizational charts, or even physical analogs such as model airplanes. Management science models are mathematical abstractions using variables and functional relationships, a symbolic form of mental model. All models are abstractions of reality, and will always leave out some elements of the reality. The key to success is to capture the essential elements in the model, along with realistic relationships among these elements. This enables experimentation to predict outcomes resulting from system changes. The model will include expected relationships of cause and effect. For instance, a model might include variables for alternative products to manufacture, such as automobile styles. These products would require resources, in terms of labor hours, tons of steel, tons of chrome, tires, and so on. If the modeler knew how each unit of each style of car used up resources, functions could be constructed to measure the hours and tons used up for any given set of variable values, each of which could be constrained to the amount available. Then if a function measuring profit for each style were available (and all of these functions were linear), an optimization model could be solved to identify the feasible combination of number of automobiles by styles identified that would have the maximum profit. Some terms involved are variables (those things you control, like numbers of automobiles by style) and parameters (constants expressing the contribution to each function, as well as constraint limits).

There are a variety of modeling types. Deterministic models have constants, making it easier to model. Linear programming is the classic deterministic modeling type. Most real decisions involve risk, in terms of parameters described by statistical distributions. Examples include waiting line systems, where arrival rates as well as service times are often probabilistic, as well as inventory models with probabilistic demand distributions. Simulation often is very useful to model systems with distributions. Uncertainty situations involve situations where statistically described distributions aren't available. These can be modeled using decision maker opinions of probabilities. Decision trees are applicable for some uncertain environments, as well as simulation. A fourth category of model type involves conflict, or competitive environments. Game theory is applicable in these environments.

Regardless of model type, a modeling process outlined earlier can be applied. Sound modeling practice should begin by assessing the overall decision environment, trying to understand the system of interacting parts and their relationships. This should lead to understanding problem components, to include who is the decision maker, what objectives are important, and what options are available. Then models can be developed by formulating the system mathematically, to include identification of variables, parameters, and functional relationships. This will require trying to express the model in some form you can solve, at the same time that you limit yourself to realistic assumptions. Once a model is constructed, using realistic data, models should be validated in that their results appear to be realistic.

It is important to remember that models are necessarily simplifications of reality. They can be made more complicated, capturing more real details. But that requires more detailed data collection, and more complicated results that may be harder to explain to managers. Modelers should always remember the assumptions included in their models, and that models are only imperfect reflections of expected reality. Nonetheless, they often enable significant improvement in business decision making.

Business analytics is a broad topic. This book focuses on modeling tools, but keep in mind that these are but one part of the overall field. Business analytics includes identifying the specific decisions that need to be made, as well as data that is needed for decision makers to make informed decisions (and that provide sound sources of input for the models we discuss). Thus understanding the business system is probably the most important step in business analytics. Then data needs to be managed. In recent years, most large organizations utilize enterprise resource planning systems heavily, but there are many forms of such systems, and they are not the only source of data. But they do provide useful means to monitor events within organizations. Sound management needs to supplement this source with others, to make sure that decision makers have a complete picture of what is going on, and the impact of their decisions needs to be monitored. If sound data is obtained, models can be highly useful.

Introduction to Business Analytics

Majid Nabavi • David L. Olson

Business analytics has grown to be a key topic in business curricula, and there is a need for stronger quantitative skills and understanding of fundamental concepts. This book is intended to present key concepts related to quantitative analysis in business. It is targeted to business students, undergraduate and graduate, taking an introductory core course.

Topics covered include knowledge management, visualization, sampling and hypothesis testing, regression (simple, multiple, and logistic), as well as optimization modeling. It concludes with a brief overview of data mining. Concepts are demonstrated with worked examples.

Majid Nabavi has published in *Quality Management Journal*, and presented research in regional and national conferences. His teaching areas include operations management, management science, database systems, and business analytics.

David L. Olson has published research in over 150 refereed journal articles, primarily on the topic of multiple objective decision making and information technology. He teaches in the management information systems, management science, and operations management areas and has authored over 20 books. He is a member of the Decision Sciences Institute, the Institute for Operations Research and Management Sciences, and the Multiple Criteria Decision Making Society. He was named the Raymond E. Miles Distinguished Scholar award for 2002, and was a James C. and Rhonda Seacrest Fellow from 2005 to 2006. He was named Best Enterprise Information Systems Educator by IFIP in 2006.

Big Data and Business Analytics Collection

Mark Ferguson, *Editor*

ISBN: 978-1-94944-327-1

90000

9 781949 443271

BEP BUSINESS EXPERT PRESS