PREDICTIVE ANALYTICS, DATA MINING and BIG DATA

MYTHS, MISCONCEPTIONS and METHODS

STEVEN FINLAY

BUSINESS IN THE DIGITAL ECONOMY

Contents

Figures and Tables

Figures

Tables

Acknowledgments

First and foremost I would like to thank my wife Samantha and my parents Paul and Ann for their support, comments and proofreading services. I would also like to thank my friend Tracy Moore for providing many useful comments and suggestions on early drafts of the manuscript. Thanks also to the staff of the Management Science Department at Lancaster University in the UK for providing access to the university facilities, which proved invaluable to my writing and research. I am also grateful to the members of the UK Government Operational Research Service (GORS), and in particular my former colleagues in Manchester and Liverpool, for many hours spent chewing over the finer points of predictive analytics, Big Data and life in general during the writing of the book.

chapter 1 Introduction

Retailers, banks, governments, social networking sites, credit reference agencies and telecoms companies, amongst others, hold vast amounts of information about us. They know where we live, what we spend our money on, who our friends and family are, our likes and dislikes, our lifestyles and our opinions. Every year the amount of electronic information about us grows as we increasingly use internet services, social media and smart devices to move more and more of our lives into the online environment.

Until the early 2000s the primary source of individual (consumer) data was the electronic footprints we left behind as we moved through life, such as credit card transactions, online purchases and requests for insurance quotations. This information is required to generate bills, keep accounts up to date, and to provide an audit of the transactions that have occurred between service providers and their customers. In recent years organizations have become increasingly interested in the spaces between our transactions and the paths that led us to the decisions that we made. As we do more things electronically, information that gives insights about our thought processes and the influences that led us to engage in one activity rather than another has become available. A retailer can gain an understanding of why we purchased their product rather than a rival's by examining what route we took before we bought it – what websites did we visit? What other products did we consider? Which reviews did we consult? Similarly, social media provides all sorts of information about ourselves (what we think, who we talk to and what we talk about), and our phones and other devices provide information about where we are and where we've been.

All this information about people is incredibly useful for all sorts of different reasons, but one application in particular is to predict future behavior. By using information about people's lifestyles, movements and past behaviors, organizations can predict what they are likely to do, when they will do it and where that activity will occur. They then use these predictions to tailor how they interact with people. Their reason for doing this is to influence people's behavior, in order to maximize the value of the relationships that they have with them.

In this book I explain how predictive analytics is used to forecast what people are likely to do and how those forecasts are used to decide how to treat people. If your organization uses predictive analytics; if you are wondering whether predictive analytics could improve what you do; or if you want to find out more about how predictive models are constructed and used in practical real-world environments, then this is the book for you.

1.1 What are data mining and predictive analytics?

By the 1980s many organizations found themselves with customer databases that had grown to the point where the amount of data they held had become too large for humans to be able to analyze it on their own. The term "data mining" was coined to describe a range of automated techniques that could be applied to interrogate these databases and make inferences about what the data meant. If you want a concise definition of data mining, then "The analysis of large and complex data sets" is a good place to start.

Many of the tools used to perform data mining are standard statistical methods that have been around for decades, such as linear regression and clustering. However, data mining also includes a wide range of other techniques for analyzing data that grew out of research into artificial intelligence (machine learning), evolutionary computing and game theory.

Data mining is a very broad topic, used for all sorts of things. Detecting patterns in satellite data, anticipating stock price movements, face recognition and forecasting traffic congestion are just a few examples of where data mining is routinely applied. However, the most prolific use of data mining is to identify relationships in data that give an insight into individual preferences, and most importantly, what someone is likely to do in a given scenario.

This is important because if an organization knows what someone is likely to do, then it can tailor its response in order to maximize its own objectives. For commercial organizations the objective is usually to maximize profit.

However, government and other non-profit organizations also have reasons for wanting to know how people are going to behave and then taking action to change or prevent it. For example, tax authorities want to predict who is unlikely to file their tax return correctly, and hence target those individuals for action by tax inspectors. Likewise, political parties want to identify floating voters and then nudge them, using individually tailored communications, to vote for them. Sometime in the mid-2000s the term "predictive analytics" became synonymous with the use of data mining to develop tools to predict the behavior of individuals (or other entities, such as limited companies). Predictive analytics is therefore just a term used to describe the application of data mining to this type of problem.

Predictive analytics is not new. One of the earliest applications was credit scoring,[1] which was first used by the mail order industry in the 1950s to decide who to give credit to. By the mid-1980s credit scoring had become the primary decision-making tool across the financial services industry. When someone applies to borrow money (to take out a loan, a credit card, a mortgage and so on), the lender has to decide whether or not they think that person will repay what they borrow. A lender will only lend to someone if they believe they are creditworthy. At one time all such decisions were made by human underwriters, who reviewed each loan application and made a decision based on their expert opinion. These days, almost all such decisions are made automatically using predictive model(s) that sit within an organization's application processing system.

To construct a credit scoring model, predictive analytics is used to analyze data from thousands of historic loan agreements to identify what characteristics of borrowers were indicative of them being "good" customers who repaid their loans or "bad" customers who defaulted. The relationships that are identified are encapsulated by the model. Having used predictive analytics to construct a model, one can then use the model to make predictions about the future repayment behavior of new loan applicants. If you live in the USA, you have probably come across FICO scores, developed by the FICO Corporation (formerly Fair Isaac Corporation), which are used by many lending institutions to assess applications for credit. Typically, FICO scores range from around 300 to about 850.[2] The higher your score the more creditworthy you are. Similar scores are used by organizations the world over. An example of a credit scoring model (sometimes referred to as a credit scorecard) is shown in Figure 1.1.

To calculate your credit score from the model in Figure 1.1 you start with the constant score of 670. You then go through the scorecard one characteristic at a time, adding or subtracting the points that apply to you,[3] so, if your

Constant	670		
Employment status		**Outstanding mortgage**	
Full-time	28	<$40,000	11
Part-time	7	$40,001–$60,000	0
Homemaker	0	$60,001–$100,000	–9
Retired	15	$100,000–$150,000	–12
Student	–8	$150,000–$250,000	–16
Unemployed	–42	> $250,000	–19
		Not a home owner	0
Time in current employment			
Not in full or part-time employment	0	**Number of credit cards**	
<1 year	–25	0	–17
1–3 years	–10	1–3	12
1–10 years	0	4–8	0
> 10 year	31	9+	–18
Residential status		**Number of days currently past due on existing credit agreements**	
Home owner	26		
Renting	0	0 (All accounts up to date)	14
Living with parent	0	1–30 days past due	0
		31–60 days past due	–29
		>60 days past due	–41
Loan amount requested as proportion of annual income			
<10%	43	**Declared bankrupt within the last 5 years?**	
10%–25%	22	Yes	–50
26%–60%	0	No	9
> 60%	–28	Unknown	0

FIGURE 1.1 Loan application model

employment status is full-time you add 28 points to get 698. Then, if your time in current employment is say, two years, you subtract 10 points to get 688. If your residential status is Home Owner you then add 26 points to get 714, and so on.

What does the score mean? For a credit scoring model the higher the score the more likely you are to repay the loan. The lower the score the more likely you are to default, resulting in a loss for the lender. To establish the relationship between score and behavior a sample of several thousand completed loan agreements where the repayment behavior is already known is required. The credit scores for these agreements are then calculated and the results used to generate a score distribution as shown in Figure 1.2.

The score distribution shows the relationship between people's credit score and the odds of them defaulting. At a score of 500 the odds are 1:1. This

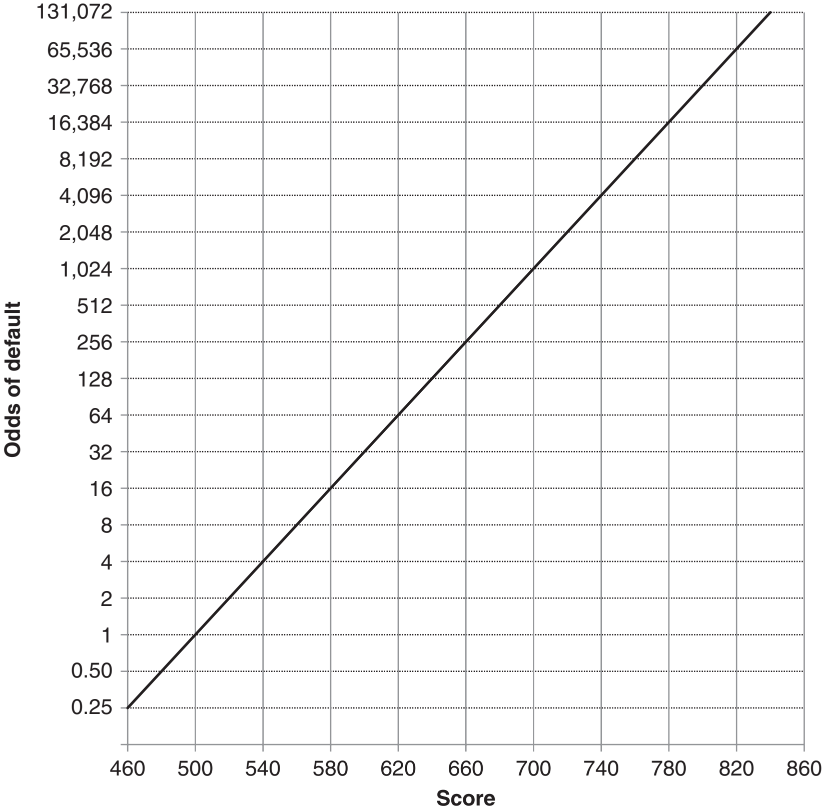


FIGURE 1.2 **Score distribution**

means that on average half of those who score 500 will default if they are granted a loan. Similarly, for those scoring 620 the odds are 64:1; i.e. if you take 65 borrowers that score 620, the expectation is that 64 will repay what they borrow, but one will not.

To make use of the score distribution in Figure 1.2 you need to have a view about the profitability of loan customers. Let's assume that we have done some analysis of all loan agreements that completed in the last 12 months. This tells us that the average profit from each good loan customer who repaid their loan was $500, but the average loss when someone defaulted was $8,000. From these figures it is possible to work out that we will only make money if there are at least 16 good customers for every one that defaults ($8,000/$500 = 16). This translates into a business decision to offer a customer a loan only if the odds of them being good are more than 16:1. You can see from the score distribution graph that this equates to a *cut-off score* of 580.

Therefore, we should only grant loans to applicants who score more than 580 and decline anything that scores 580 or less. So given the model in Figure 1.1, do you think that you would get a loan?

An absolutely fundamental thing to understand about a predictive model like this is that we are talking about probability, not certainty. Just like a human decision maker, no model of consumer behavior gets it right every time. We are making a prediction, not staring into a crystal ball. Whatever score you get does not determine precisely what you will do. Scoring 800 doesn't mean you won't default, only that your chance of defaulting is very low (1 in 32,768 to be precise). Likewise, for people scoring 560 the expectation is that eight out of every nine will repay – still pretty good odds, but this isn't a pure enough pot of good customers to lend profitability based on an average profit of $500 and an average loss of $8,000. It's worth pointing out that although the credit industry talks about people in terms of being "creditworthy" or "uncreditworthy," in reality most of those deemed uncreditworthy would actually repay a loan if they were granted one.

Some other important things to remember when talking about credit scoring models (and predictive models in general):

- **Not all models adopt the same scale.** A score of 800 for one lender does not mean the same thing as 800 with another.
- **Some models are better than others.** One model may predict your odds of default to be 20:1 while another estimates it to be 50:1. How good a model is at predicting behavior depends on a range of factors, in particular the amount and quality of the data used to construct the model, and the type of model constructed. (Scorecards are a very popular type of model, but there are many other types, such as decision trees, expert systems and neural networks.)
- **Predictions and decisions are not the same thing.** Two lenders may use the same predictive model to calculate the same credit score for someone, but each has a different view of creditworthiness. Odds of 10:1 may be deemed good enough to grant loans by one lender, but another won't advance funds to anyone unless the odds are more than 15:1.

1.2 How good are models at predicting behavior?

In one sense, most predictive models are quite poor at predicting how someone is going to behave. To illustrate this, let's think about a traditional paper-based mail shot. Although in decline, mail shots remain a popular tool employed by marketing professionals to promote products and services to consumers. Consider an insurance company with a marketing strategy that involves sending mail shots to people offering them a really good deal on life insurance. The company uses a response model to predict who is most likely to want life insurance, and these people are mailed.

If the model is a really good one, then the company might be able to identify people with a 1 in 10 chance of taking up the offer – 10 out of every 100 people who are mailed respond. To put it another way, the model will get it right only 10% of the time and get it wrong 90% of the time. That's a pretty high failure rate! However, what you need to consider is what would happen without the model. If you select people from the phone book at random, then a response rate of around 1% is fairly typical for a mail shot of this type. If you look at it this way, then the model is ten times better than a purely random approach – which is not bad at all.

In a lot of ways we are in quite a good place when it comes to predictive models. In many organizations across many industries, predictive models are generating useful predictions and are being used to significantly enhance what those organizations are doing. There is also a rich seam of new applications to which predictive analytics can be applied. However, most models are far from perfect, and there is lots of scope for improvement. In recent years, there have been some improvements in the algorithms that generate predictive models, but these improvements are relatively small compared to the benefits of having more data, better quality data and analyzing this data more effectively. This is the main reason why Big Data is considered such a prize for those organizations that can utilize it.

1.3 What are the benefits of predictive models?

In many walks of life the traditional approach to decision making is for experts in that field to make decisions based on their expert opinion. Continuing with our credit scoring example, there is no reason why local bank managers can't make lending decisions about their customers (which is what they used to do in the days before credit scoring) – one could argue that this would add that personal touch, and an experienced bank manager should be better able to assess the creditworthiness of their customers than some impersonal credit scoring system based at head office. So why use predictive models?

One benefit is speed. When predictive models are used as part of an automated decision-making system, millions of customers can be evaluated and dealt with in just a few seconds. If a bank wants to produce a list of credit card customers who might also be good for a car loan, a predictive model allows this to be undertaken quickly and at almost zero cost. Trawling through all the bank's credit card customers manually to find the good prospects would be completely impractical. Similarly, such systems allow decisions to be made in real time while the customer is on the phone, in branch or online.

A second major benefit of using predictive models is that they generally make better forecasts than their human counterparts. How much better depends on the problem at hand and can be difficult to quantify. However, in my experience, I would expect a well-implemented decision-making system, based on predictive analytics, to make decisions that are about 20–30% more accurate than their human counterparts. In our credit scoring example this translates into granting 20–30% fewer loans to customers who would have defaulted or 20–30% more loans to good customers who will repay, depending upon how one decides to use the model. To put this in terms of raw bottom line benefit, if a bank writes off $500m in bad loans every year, then a reasonable expectation is that this could be reduced by at least $100m, if not more, by using predictive analytics. If we are talking about a marketing department spending $20m on direct marketing to recruit 300,000 new customers each year, then by adopting predictive analytics one would expect to spend about $5m less to recruit the same number of customers. Alternatively, they could expect to recruit about 75,000 more customers for the same $20m spend.

A third benefit is consistency. A given predictive model will always generate the same prediction when presented with the same data. This isn't the case with human decision makers. There is lots of evidence that even the most competent expert will come to very different conclusions and make different decision about something depending on their mood, the time of day, whether they are hungry or not and a host of other factors.[4] Predictive models are simply not influenced by such things. This leads on to questions about the bias that some people display (consciously or unconsciously) against people because of their gender, race, religion age, sexual orientation and so on. This is not to say that predictive models don't display bias towards one group or another, but that where bias exists it is based on clear statistical evidence. Many types of predictive model, such as the scorecard in Figure 1.1, are also explicable. It's easy to understand how someone got the score that they did, and hence why they did or did not get a loan. Working out why a human expert came to a particular decision is not always so easy, especially if it was based on a hunch. Even if the decision maker keeps detailed notes, interpreting what they meant isn't always easy after the event.

Is it important for a predictive model to be explicable? The answer very much depends on what you are using the model for. In some countries, if a customer has their application for credit declined it is a legal requirement to give them an objective reason for the decision. This is one reason why simple models such as those in Figure 1.1 are the norm in credit granting. However, if you are using predictive models in the world of direct marketing, then no one needs to know why they did or didn't get a text offering them a discount on their next purchase. This means that the models can be as simple or as complex as you like (and some can be very complex indeed).

1.4 Applications of predictive analytics

Credit scoring was the first commercial application of predictive analytics (and remains one of the most popular), and by the 1980s the same methods were being applied in other areas of financial services. In their marketing departments, loan and credit card providers started developing models to identify the likelihood of response to a marketing communication, so that only those most likely to be interested in a product were targeted with an offer. This saved huge sums compared to the blanket marketing strategies that went before, and enabled individually tailored communications to be sent to each person based on the score they received. Similarly, in insurance predictive models began to be used to predict the likelihood and value of claims. These predictions were then used to set premiums.

These days, predictive models are used to predict all sorts of things within all sorts of organizations – in fact, almost anywhere where there is a large population of individuals that need decisions to be made about them. The following is just a small selection of some of the other things that predictive models are being used for today:[5]

1. Identifying people who don't pay their taxes.
2. Calculating the probability of having a stroke in the next 10 years.
3. Spotting which credit card transactions are fraudulent.
4. Selecting suspects in criminal cases.
5. Deciding which candidate to offer a job to.
6. Predicting how likely it is that a customer will become bankrupt.
7. Establishing which customers are likely to defect to a rival phone plan when their current contract is up.
8. Producing lists of people who would enjoy going on a date with you.
9. Determining what books, music and films you are likely to purchase next.

10. Predicting how much you are likely to spend at your local supermarket next week.
11. Forecasting life expectancy.
12. Estimating how much someone will spend on their credit card this year.
13. Inferring when someone is likely to be at home (so best time to call them).

The applications of predictive models in the above list fall into two groups. Those in the first group are concerned with yes/no type questions about behavior. Will someone do something or won't they? Will they carryout action A or action B? Models that predict this type of behavior are called classification models. The output of these models (the model score) is a number that represents the probability (the odds)[6] of the behavior occurring. Sometimes the score provides a direct estimate of the likelihood of behavior. For example, a score of 0.4 means the chance of someone having a heart attack in the next five years is 40% (and hence there is a 60% chance of them not having one). In other cases the score is calibrated to a given scale – perhaps 100 means the chance of you having a heart attack is the same as the population average. A score of 200 twice average, a score of 400 four times average and so on. For the scorecard in Figure 1.1, the odds of default double every 20 points – which is a similar scale to the one FICO uses in its credit scores.

All of the first nine examples in the above list can be viewed from a classification perspective (although this may not be obvious at first sight). For example, an online bookseller can build a model by analyzing the text in books that people have bought in the past to predict the books that they subsequently purchased. Once this model exists, then your past purchasing history can be put through the model to generate a score for every book on the bookseller's list. The higher the score, the more likely you are to buy each book. The retailer then markets to you the two or three books that score the most: the ones that you are most likely to be interested in buying.

The second type of predictive model relates to quantities. It's not about whether you are going to do something or not, but the magnitude of what you do. Typically, these equate to "how much" or "how long" type questions. Actuaries use predictive models to predict how long people are going to live, and hence what sort of pension they can expect. Credit card companies build value models to estimate how much revenue each customer is likely to generate. These types of models are called regression models (items 10–13 in the list). Usually, the score from a regression model provides a direct estimate of the quantity of interest. A score of 1,500 generated by a revenue model means that the customer is expected to spend $1,500. However, sometimes what one is interested in is ranking customers, rather than absolute values. The model might be constructed to generate scores in the range 1–100, representing the percentile into which customer spending falls. A score of 1 indicates that the customer is in the lowest spending percentile and a score of 100 that they are in the highest scoring percentile.

In terms of how they look, classification and regression models are very similar, but at a technical level there are subtle differences that determine how models are constructed and used. Classification models are most widely applied, but regression models are increasingly popular because they give a far more granular view of customer behavior. At one time a single credit scoring model would have been used to predict whether or not someone was likely to repay their loan, but these days lenders also create models to predict the expected loss on defaulting loans and the expected revenues from good paying ones. All three models are used in combination to make much more refined lending decisions than could be made by using a single model of loan default on its own.

1.5 Reaping the benefits, avoiding the pitfalls

An organization that implements predictive analytics well can expect to see improvements in its business processes of 20–30% or even more in some cases. However, success is by no means guaranteed. In my first job after graduation, working for a credit reference agency more than 20 years ago, I was involved in building predictive models for a number of clients. In general the projects went pretty well. I delivered good-quality predictive models and our clients were happy with the work I had done and paid accordingly. So I was pretty smug with myself as a hot shot model builder. However, on catching up with my clients months or years later, not everyone had a success story to tell. Many of the models I had developed had been implemented and were delivering real bottom line benefits, but this wasn't universally the case. Some models hadn't been implemented, or the implementation had failed for some reason.

Digging a little deeper it became apparent that it wasn't the models themselves that were at fault. Rather, it was a range of organizational and cultural issues that were the problem. There are lots of reasons why a predictive analytics project can fail, but these can usually be placed into one of three categories:

1. **Not ready for predictive analytics**. Doing something new is risky. People are often unwilling to take the leap of faith required to place trust in automated models rather than human judgment.

2. **The wrong model.** The model builder thought their customer wanted a model to predict one type of consumer behavior, but the customer actually wanted something that predicted a different behavior.
3. **Weak governance.** Implementing a predictive model sometimes requires changes to working practices. As a rule, people don't like change and won't change unless they have to. Just telling them to do something different or issuing a few memos doesn't work. Effective management and enforcement are required.

More than 20 years after I had this realization, methods for constructing predictive models and the mechanisms for implementing predictive models have evolved considerably. Yet I still frequently hear of cases where predictive analytics projects have failed, and it's usually for one of these reasons.

One thing to bear in mind is that different people have different views of what a project entails. For a data scientist working in a technical capacity, a predictive analytics project is about gathering data and then building the best (most predictive) model they can. What happens to the model once they have done their bit is of little concern. Wider issues around implementation, organizational structures and culture are way out of scope.

Sometimes this is fine. If an organization already has an analytics culture and a well-developed analytics infrastructure, then things can be highly automated and hassle-free when it comes to getting models into the business. If the marketing department is simply planning to replace one of its existing response models with a new and a better one, then all that may be involved is hitting the right button in the software to upload the new model into the production environment. However, the vast majority of organizations are not operating their analytics at this level of refinement (although many vendors will tell you that everyone else is, and you need to invest in their technology if you don't want to get left behind). In my experience, it's still typical for model building to account for no more than 10–20% of the time, effort and cost involved in a modeling project. The rest of the effort is involved in doing all the other things that are needed to get the processes in place to be able to use the model operationally.

Even in the financial services industry, where predictive models have been in use longer than anywhere else, there is a huge amount that people have to do around model audit and risk mitigation before a model to predict credit risk can be implemented.[7] What this means in practice is that if you are going to succeed with predictive analytics, you need a good team to deliver the goods. This needs to cover business process, IT, data and organizational culture, with good project management to oversee the lot. Occasionally, a really top class data scientist can take on all of these roles and do everything from the gathering initial requirements through to training staff in how to use the model, but these multi-skilled individuals are rare. More often than not, delivery of analytical solutions is a team effort, requiring input from people from across several different business areas to make it a success.

1.6 What is Big Data?

Large and complex data sets have existed for decades. In one sense Big Data is nothing new, and for some in the industry all the hype around Big Data came as a bit of a surprise. "The emperor's new clothes!" they cried. However, by the early 2010s "Big Data" had become the popular catch-all phrase to describe databases that are not just large, but enormous and complex. There isn't a universally agreed definition of "Big Data," but the features of Big Data[8] that are considered important are:

- **Volume.** Any database that is too large to be comfortably managed on an average PC/laptop/server can be considered Big Data. At the time of writing, Big Data is generally taken to be a database that contains more than a terabyte (1,000 gigabytes) of data.[9] Some Big Data sources contain petabytes of data (1 petabyte = 1,000 terabytes).
- **Variety.** Big Data contains many different types of structured and unstructured data. Structured data is tidy and well defined and can usually be represented as numbers or categories: for example your income, age, gender and marital status. Unstructured data is not well defined. It is often textual and difficult to categorize: for example e-mails, blogs, web pages and transcripts of phone conversations.
- **Volatility (velocity).** Some types of data are relatively static, such as someone's place of birth, gender, social security number and nationality. Other data changes occasionally/slowly, such as one's address, your employer or the number of children that you have. At the other extreme data is changing all the time: for example what music you are listening to right now, the speed you are driving and your heart rate. Big Data is often volatile.
- **Multi-sourced.** Some Big Data sources are generated entirely from an organization's internal systems. This means they have control over its structure and format. However, Big Data often includes external data such as credit reports, census information, GPS data and web pages, and

organizations have little control over how it's supplied and formatted. This introduces additional issues around data quality, privacy and security, over and above what is required from internally sourced data.

This is the sort of definition you usually hear when describing Big Data,[10] but it's difficult to quantify. It doesn't give you a clear dividing line between normal data and Big Data, and therefore many people find it vague and confusing. Using this definition you can't say that a data set is Big Data simply because it contains a particular type of data or a certain amount of data. There is also the onward march of technology to consider. What was Big Data yesterday is not Big Data today, and what's Big Data today may not be Big Data tomorrow – the goal posts are always shifting. There is also a degree of context involved when one talks about Big Data. What constitutes Big Data for a small company, with a few tens of thousands of customers, may be very different from the view held by a large multinational with tens of millions of customers.

Rather than getting hung up on a precise definition of Big Data, an alternative perspective is to view Big Data as a philosophy about how to deal with data, rather than how much data is available or what it contains. The four tenets of this philosophy are:

1. Seek.
2. Store.
3. Analyze.
4. Act.

You proactively search for and obtain new data: you bring all your data together and analyze it to produce insights about what people have done, what they are doing and what they are likely to do in the future. This in turn informs your decision making and what actions to take. A Big Data philosophy is about taking a holistic view of the data available to you and getting the best out of what you have. If an organization is doing this then it doesn't really matter if it has just a few megabytes or many petabytes of data; if the data is structured or unstructured; or where it comes from. From a technology perspective one seeks out IT solutions that deliver the required storage and analytical capability. In some situations this might mean using newer technologies such as Hadoop or Storm, but a traditional relational database solution is often sufficient (or even superior[11]) and should not be ruled out.

So you can view Big Data from a number of perspectives, but for the rest of this book we'll keep things simple and adopt a fairly laid-back definition of

Big Data as: "A very large amount of varied data." So to give a few examples: A government census containing a few dozen items of information about each of that country's 250 million citizens, such as social security number, gender, marital status, income and number of children is a large amount of data, but probably not Big Data. Likewise, the text from 50,000,000 internet pages isn't Big Data either. However, an organization's database of all the information it has gathered about its three million customers in the last five years (purchase details, billing information, e-mails, server logs, texts, transcripts of phone calls, complaint letters, notes taken by staff in branch, credit reports, GPS data and so on) is Big Data.

The first myth I want to dispel is that you need a huge amount of data to build a predictive model. A couple of thousand customer records and a few dozen choice pieces of information about those customers are more than enough, and many useful models have been built using less data than this. However, the more data you have about people, the more predictive your models will be.[12] Big Data has attracted so much interest in recent years because it goes beyond the traditional data sources that people have used in data mining and predictive analytics in the past. In particular, when people talk about Big Data, what they often mean is:

- **Textual data.** This comes from letters, phone transcripts, e-mails, webpages, tweets and so on. This type of data is unstructured and therefore needs a lot of processing power to analyze it.
- **Machine-generated data.** For example, GPS data from people's phones, web logs that track internet usage and telematic devices fitted to cars. Machine-generated data is generally well structured and easy to analyze, but there is a lot of it.
- **Network data.** This is information about people's family, friends and other associates. In some contexts what is important is the structure of the network to which an individual belongs – how many people are in the network, who is at the center of the network and so on. In other contexts the network is a mechanism for inferring things about someone, based on the features of other people in their network.

It used to be the case that the prime source of data for all sorts of predictive models was well-structured internal data sources, possibly augmented by information from a credit reference agency or database marketing company, but these days Big Data that combines traditional data sources with these new types of data is seen as the frontier in terms of consumer information. The problem is that there is so much different and varied data around – so much so that it is becoming increasingly difficult to analyze it all. There is something of an arms race going on between the IT community, who are continuously developing their hardware and software to obtain and store more and more differing and diverse data, and the analytical fraternity, who are trying to find better and more efficient ways to squeeze useful insights from all the data that their IT colleagues have gathered.

1.7 How much value does Big Data add?

When long-established users of predictive analytics, such as banks, insurers and retailers, ask about the value of Big Data, what they really want to know is: What uplift to our predictive models will Big Data provide, over and above the data we use already? I would not go so far as to say that there aren't other applications for Big Data – there are several[13] – but using it to develop predictive models is what people often have in mind. One feature of Big Data is that most of it has a very low information density, making it very difficult to extract useful customer insights from it. A huge proportion of the Big Data out there is absolutely useless when it comes to forecasting consumer behavior. You have to work pretty hard at finding the useful bits that will improve the accuracy of your predictive models – and this is why you need big computers with lots of storage, and clever algorithms, to find the important stuff amongst the chaff.

The way I like to think about this – as have many others – is by an analogy with gold mining.[14] Most industrial activity involved in gold extraction occurs in mines where there are rich seams full of nice big nuggets. Yet it has been estimated that there are thousands of tons of gold dissolved in the world's oceans, possibly more than the entire amount that has been mined across all of human history.[15] However, the gold is very diffuse amongst all that seawater, and no one has yet found a cost-effective way of extracting it. Traditional customer databases are analogous to the gold mine, while Big Data is the ocean. There is lots of useful information floating around in all that data, but it's very sparse compared to traditional data sources and it can be very expensive to get at it.

While we are on the subject of gold, it's worth remembering that in gold rushes (and internet booms and the exciting world of Big Data) the people who sell the tools make a lot of money. Far more strike it rich selling picks and shovels to prospectors than do the prospectors. Likewise, there is a lot of money to be made selling Big Data solutions. Whether the buyer actually gets any benefit from them is not the primary concern of the sales people. There are lots of opportunities for Big Data, but Big Data is not the answer to all of the world's problems and it may not be right for you.

Sometimes the benefits of Big Data can be too small to justify the expense. In banking, for example, the potential for new Big Data sources to improve the predictive ability of credit scoring models is fairly small, over and above the data already available. This is because the key driver of credit risk is past behavior, and the banks have ready access to people's credit reports, plus a wealth of other data, supplied in a nice neat format by Credit Reference Agencies such as Equifax, Experian and TransUnion. At the other end of the spectrum, if you are a marketing team trying to identify people who might be interested in your products and you have nothing to go on, then externally sourced Big Data can provide a lot of value. Being able to trawl the internet, server logs, social network sites, tweets and so on to find out who is talking about what, or what people's friends and family are buying (social network analysis, as discussed in Chapter 9) has immense value over and above the very small amount you would otherwise know about people.

For those already using predictive analytics, one view is that Big Data is very much the icing on the cake once you have good IT systems and good analytics in place – it's a process of evolution not revolution. You can also "bolt on" new technologies such as Hadoop to what you already have. It's not a question of "either/or." However, if your internal data systems are a mess, you don't store as much data as you could and you don't have a strong analytics culture, then Big Data solutions are probably not the next step for you. It's very much a case of trying to run before you can walk. Some organizations have made the leap directly to a Big Data/Analytics culture, but that's rare. It's a high-risk strategy and one that requires commitment, extensive organizational restructuring and a lot of expense. My recommendation is that you should concentrate on getting your own house in order, and making better use of the easy to access data you already have, before moving on to more complex solutions that encompass a wider Big Data philosophy. To put it another way, unless your IT and analytics are already pretty slick, you will get far more bang for your buck from incremental improvements to your current systems, compared to implementing a whole new suite of dedicated hardware and software specifically for handling Big Data.

In terms of the percentage uplift that Big Data provides, that's something of an open question, and is very dependent upon the type of predictive models you want to build and how much data you already make use of. All sorts of figures get bandied about, and as with predictive analytics, you'll only tend to hear about the success stories rather than the failures. Therefore expectations tend to become somewhat over-inflated.[16] Using Big Data to enhance your existing predictive modeling capability is a second-order effect. This means that if you adopt a Big Data approach then you can expect your models to improve, but the improvement won't be anything like as much as when predictive models were implemented for the first time.

My view is that if you already have good data and analytics, and you implement a Big Data Strategy in the right way, then you may see a 4–5% uplift in the performance of your predictive models.[17,18] However, it really depends on the amount of good quality data you already have and what new information your Big Data sources are bringing to the party. If you don't currently have much customer data, and Big Data gives you the ability to predict customer behavior where this wasn't an option before, then you could be looking at benefits of significantly more than 10%. However, if you already have loads of well-organized, accurate and nicely formatted data, then expect to see more modest returns on your investment.

Another perspective, and the one I adhere too, is that the biggest benefits of Big Data have little to do with enhancing existing models in well-run data-rich organizations. Sure, there are some benefits to be had, but the greatest opportunities for Big Data are where it is making new forms of customer prediction viable. One example is where police forces are using huge databases of past crimes to predict the neighborhoods where a crime is likely to be committed in the next few hours. They can then concentrate their resources in those areas.[19] Another area of huge potential is preventative healthcare. Most existing healthcare systems are reactive: they treat you when you are already ill. Combining predictive analytics with Big Data makes it more viable to shift the emphasis to prevention. It becomes possible to predict how likely each citizen is to develop certain conditions and intervene before the illness becomes apparent. This has the potential to add years to average life expectancy.

Marketing is another area where Big Data is proving its worth. For example, by combining information about your movements, gathered from your cell phone, with supermarket data about what type of food you like to buy, you can be targeted with promotional offers for restaurants in the city you are traveling to before you even get there. Another marketing application is to use real-time information about electricity and gas usage to forecast when someone is likely to be at home, and therefore a good time to contact them.

These applications of predictive analytics were little more than science fiction just a few years ago, but this is where the frontier of Big Data and predictive analytics currently lies.

1.8 The rest of the book

I hope this introductory chapter has given you an insight into what Big Data and predictive analytics are, what (one type) of predictive model looks like and how models can be used to help organizations achieve their objectives.

With regard to the rest of the book, Chapters 2–5 are very much about the application and usage of predictive models. Chapter 2 explains how models are used in organizations to drive decision making, and Chapter 3 discusses the analytics culture that needs to be in place if the benefits of predictive analytics are to be realized. In Chapter 4 the focus is on data, in particular the types of data that an organization needs to build good quality predictive models. We also consider the relative value of different types of data, and what data sources an organization should focus on given its current capabilities. There is after all no point getting all excited about Big Data if your existing customer databases are in poor shape. In Chapter 5, we consider ethical and legal issues associated with personal data, and the use of that data within automated decision-making systems built around predictive models.

In the second part of the book, in Chapters 6–10, the focus is more about the process of developing and implementing predictive models. Chapter 6 explains, compares and contrasts the popular types of predictive model that are in use today. This includes linear models, decision trees, expert systems, support vector machines and neural networks. The current trend of generating forecasts of consumer behavior by combining the outputs of several different predictive models (ensemble systems), is also covered.

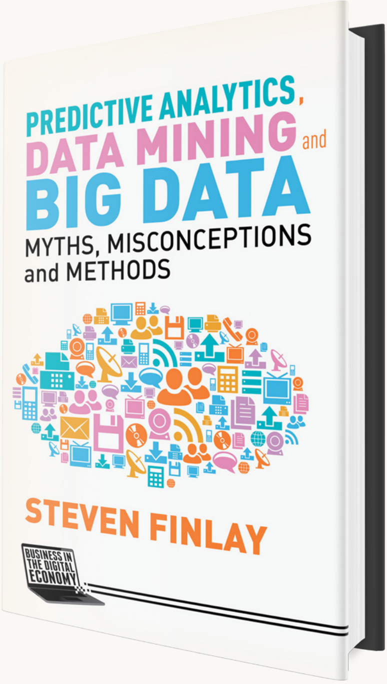
Chapter 6 is the most technical chapter in the book, but I have attempted to explain everything in a non-technical "formula-free" way without mathematics. However, if it's not your cup of tea, then you can skip this chapter without significant risk of getting lost in subsequent chapters.

Chapters 7 and 8 cover the end-to-end predictive analytics process – what needs to be done to build good quality predictive models and get them implemented within an organization. Chapter 7 discusses each stage of an analytical project, starting with project planning and going right through to implementation within the business and the post-implementation tasks required to ensure that the model continues to generate good quality predictions. Chapter 8 then describes how you go about building a predictive model once you have planned out what you are going to do.

In Chapter 9 we consider two more recent types of data analysis techniques that can be used to enhance the accuracy of predictive models by extracting some of the information that is hidden within Big Data. The first of these is Text Mining (Text Analytics). Text Mining is the art of extracting useful information from speech and text, such as web pages, emails, transcripts of phone conversations and so on. The other method discussed in Chapter 9 is Social Network Analysis. This is about the relationships people have with each other, and how information about our associates can improve the accuracy of predictive models. The final chapter, Chapter10, discusses some of the IT and software issues relating to predictive analytics and Big Data.

PREDICTIVE ANALYTICS

The Essential Guide to Improving Operational Efficiency


PREDICTIVE ANALYTICS, DATA MINING and BIG DATA
MYTHS, MISCONCEPTIONS and METHODS
STEVEN FINLAY
BUSINESS IN THE DIGITAL ECONOMY

9781137379276 I July 2014
~~£29.99~~ £20.99 I ~~$45~~ $31.50 I Hardback

Predictive analytics, data mining and big data are key topics for organizations who want to leverage the ever increasing amounts of data that they hold about people.

This easy to read, in-depth guide provides readers with a solid understanding of predictive analytics, and how it should be applied to improve business decision making and operational efficiency. This includes how to avoid the pitfalls and dangers of introducing predictive analytics to a business area for the first time, legal, ethical and cultural issues that need to be considered, and a contextual road map for developing solutions that deliver real benefits to organizations. This how-to-guide will help managers to make the most of these technologies in their business area.

About the Author

STEVEN FINLAY is one of the UK's leading experts on predictive analytics and its application within Big Data environments. He has extensive experience of developing predictive analytics solutions within Financial Services, Retailing and Government organisations. Steven is currently Head of Analytics at HML, the UK's largest provider of mortgage administration services. Previously he has worked as a data scientist, consultant and project manager for a variety of organizations in both the public and private sectors. Steven has a PhD in predictive analytics and is an Honorary Research Fellow in the Management Science Department at Lancaster University in the UK.

SAVE 30%
Enter PM15THIRTY when
ordering your copy at palgrave.com

palgrave macmillan

PROFESSIONAL BUSINESS

@PalgraveBiz
Palgrave Professional Business Books