



ELSEVIER

Available online at www.sciencedirect.com



International Journal of Forecasting 25 (2009) 441–451

*international journal
of forecasting*

www.elsevier.com/locate/ijforecast

Mining the past to determine the future: Problems and possibilities

David J. Hand*

*Department of Mathematics, Imperial College, London, United Kingdom
Institute for Mathematical Sciences, Imperial College, London, United Kingdom*

Abstract

Technological advances mean that vast data sets are increasingly common. Such data sets provide us with unparalleled opportunities for modelling and predicting the likely outcome of future events. However, such data sets may also bring with them new challenges and difficulties. An awareness of these, and of the weaknesses as well as the possibilities of these large data sets, is necessary if useful forecasts are to be made. This paper looks at some of these difficulties, using illustrations with applications from various areas.

© 2008 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Empirical models; Iconic models; Data mining; Model search; Large datasets; Selection bias

It is utterly implausible that a mathematical formula should make the future known to us, and those who think it can would once have believed in witchcraft.

Jacob Bernoulli, in *Ars Conjectandi*, 1713

1. Introduction

Modern data capture technologies and the capacity for data storage mean that we are experiencing a data deluge. This brings with it both opportunities and challenges. The opportunities arise from the

possibility of discerning structures and patterns which would be undetectable with data sets with fewer points or which did not include such a range of variables. The challenges include those of searching through such vast data sets, as well as issues of data quality and apparent structure arising by chance. Such issues are discussed by Hand, Blunt, Kelly and Adams (2000).

Forecasting has always been an important statistical problem — indeed, it certainly predates the development of formal data analytic tools. But with the development of formal analytics, highly sophisticated forecasting methods have been developed, with particular tools created for the unique problems of different kinds of domain.

When the two areas come together — forecasting based on large masses of data and using the rapid development tools of data mining — new

DOI of original article: [10.1016/j.ijforecast.2008.11.001](https://doi.org/10.1016/j.ijforecast.2008.11.001).

* Corresponding address: Imperial College, Department of Mathematics, South Kensington Campus, London SW7 2AZ, United Kingdom.

E-mail address: d.j.hand@imperial.ac.uk.

opportunities are created. But, as with data mining in general, such opportunities do not come without their caveats. The careless use of any sophisticated tool can lead to misleading conclusions, and data mining is no exception. It is my view that these dangers have been largely overlooked by the data mining community, and, now that the discipline is firmly established, they need to be addressed. In this paper I briefly summarise high level notions of forecasting and data mining, and then look at some of these dangers. I illustrate these points using examples from various domains, though most come from the personal financial services sector, partly because I have considerable experience in that area, and partly because many of the dangers are particularly apparent in that area.

2. Forecasting

Economists joke that steering the economy is like steering a car by looking through the rear view mirror. Of course, one would never steer a car like that. To steer a car, one looks ahead, noting that one is approaching a bend in the road, that there is another vehicle bearing down on one, and that there is a cyclist just ahead on the near side. That is, in steering a car, one sees that certain things lie ahead, which will have to be taken into account. The presumption in this joke is that in steering the economy one cannot see what lies ahead, but, instead, has to try to predict it based on an analysis of past data.

In such a retrospective analysis, one examines configurations of incidents from the past, seeking arrangements which are similar to those of the present, so that one can extrapolate from these past incidents through the present to the future. Sophisticated extrapolations also take into account the uncertainties involved, giving distributions or confidence intervals for likely future values.

The fact is, however, that in steering a car one is making exactly the same kind of retrospective analysis. One observes the car ahead, and, *based on one's previous experience with approaching vehicles*, assumes that the vehicle will continue to proceed, in a relatively uniform manner, on the correct side of the road.

The key to this, in the cases of both the economy and the car, is that one's predictions, one's forecasts, are based on assumptions of continuity with one's past

experience. If, in many similar situations in the past, almost all had been followed by a particular event, then one would have considerable confidence that the same thing would happen the next time: the sun rising tomorrow is the classic example. The trick in all of this is quantifying the degree of continuity, and, in a sense, that is what all forecasting is about.

The desire to forecast is universal — one of those things we all wish we could do is know the future. Forecasting has several aspects. One is defining the degree of similarity between the present and the past. Another is determining the range and variability of events which followed these similar past events, and a third is deciding whether one understands enough about the underlying process to adopt a particular model form.

Forecasting also has its limitations. Firstly, there are chaotic limitations. These are fundamental in the sense that they tell us that no matter how much we know about the past and the present, and no matter how accurately we know it, there comes a point in the future at which the minuscule inaccuracies in our present knowledge will have been amplified to render our forecasts wide of the mark. This is nicely illustrated by weather forecasting, where, thanks to vast computational resources and huge statistical models, we can now forecast reasonably accurately perhaps five days ahead, but where extending this significantly further requires dramatically more extensive data and greater computer power.

Secondly, there are stochastic limitations. These are the sudden, unpredicted and unpredictable jolts to the system which are often caused by external agencies, or perhaps by inadequacies in the model. A nice recent example of that is the current global financial crisis. I have been asked: could we have seen it coming? The short answer, of course, is that we did: there are many economic forecasters, and at least as many forecasts. Some of these were sufficiently confident of the danger to act on it (some hedge funds did very well out of it). If we combine data mining with forecasting we can always find someone who (on looking back) gave the right forecast. This is the basis for the sure-fire way of making money as a stock market tipster by making a series of multiple different forecasts, and eventually selecting just those potential customers to whom you gave a series which happened to turn out to be correct. It also illustrates the difficulties of making

inferences in data mining, when huge data sets and numbers of data configurations are involved.

3. Data mining

The preface of my book *Principles of Data Mining* (Hand, Mannila, & Smyth, 2001) opened by defining data mining as ‘the science of extracting useful information from large data sets or databases’. I think that this brief definition is sufficiently broad that it will be non-controversial. However, the opening chapter of the book then included the more detailed definition: ‘the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.’ A comparison of this definition with past and current data mining practice immediately reveals something important: data mining, as a discipline, is developing and changing. Perhaps this is hardly surprising. The discipline is a young one — necessarily so, since it is a child of the computer age — and young things do grow and develop.

In its earliest usages, the term data mining was typically used in a derogatory sense (along with data ‘snooping’, ‘fishing’, ‘trawling’, and so on). It meant the examination of data sets from a large number of angles, fitting a great many models, or looking at a great many subsets of the data. The data set may not have been very large by modern standards (we are talking early days of the computer), but the number of possible data examinations which could be applied was essentially unlimited. The derogatory implication arises from the truism that, in any given data set, if you look hard enough, you are bound to find apparently unusual data configurations.

I think that the two aspects of this early perspective on data mining encapsulated the typical statistician’s perspective around, say, the 1980s — firstly, that it involved an extensive model search, and secondly, the derogatory implication. The interpretation of data mining as being primarily about extensive model searching continues to be pertinent. For example, Hoover and Perez (1999) entitled their paper on model search: ‘*Data mining reconsidered: encompassing and the general-to-specific approach to specification search*’. However, the derogatory implication has died away as the notion of data mining as being

about extracting useful knowledge from large data sets has come to the fore. This is probably partly as a consequence of the manifest need to do this in many situations. In addition, however, the further perspective that data mining is concerned with seeking information in large data sets has become more important nowadays, concurrently with the growth in numbers, and indeed sizes, of large data sets.

It is my personal view also that the changing population of researchers involved in data mining was a key cause of the improved regard in which data mining is held. In particular, as computers developed, so computer scientists, with backgrounds in database technology and related areas, gradually became more concerned with the analysis of data. Indeed, entire computational disciplines concerned with data analysis grew up, such as machine learning and pattern recognition. My personal (entirely subjective gross generalisation) observation is that (on average!) computer scientists are less conservative than statisticians, so that, where a statistician might prefer to err on the side of caution in sifting through data, a computer scientist might give it a whirl. Furthermore, with a background more solidly in data storage and the properties of existing data sets (through work in databases), in the earlier days of data mining computer scientists were less concerned with notions of inference — that is, of generalising from the configurations found in the database to patterns in data sets yet to be collected. This would have made them less aware of the role of chance and random variation in producing apparently interesting data configurations. For example, if one’s main concern is with the data in the personnel database, it is the people actually described there in which one is interested. Chance, and the personnel characteristics we would observe if we had had another set of employees ‘drawn from the same distribution as that of the actual employees’, is an irrelevant and uninteresting question. My suspicion that, at least in the early days of data mining, computer scientists were less concerned with the inferential issues, and more concerned with the description of actual existing data sets, is not mere speculation: I took part in interdisciplinary debates on such matters in the early 1990s.

If one’s primary aim is simply to summarise or describe particular features of a given data set, then

many of the subtle difficulties and problems which arise with inference become irrelevant. However, inference is central to forecasting, so these problems now become central. I think that the fact that mining large data sets originally sprang from computational rather than statistical roots explains why data miners have so often not appreciated just how important and tough these issues are.

Data mining is also changing in other ways. In particular, the extended definition above qualifies the data sets as ‘observational’. That is, data sets were regarded not as the product of designed experimentation, but were collected (often as a by-product of some other exercise) simply by measuring the world as it is presented. It is true that most data mining is still carried out on such observational data sets. In some domains (astronomy and archaeology, for example), the nature of the domain of study renders experimentation impossible. However, in other domains (business applications and medical research, for example) experimentation is certainly possible, and we are witnessing the collection of very large data sets which have been collected as an integral part of a data mining process. A classic example of this is the experimentation work carried out by the bank Capital One, of which I say more below.

There are also differences in the way data mining is used in different applications. This is true of both of the interpretations noted above: the exploration of a huge model space and the exploration of a huge data space. In scientific applications, for example, one often finds quite sophisticated techniques being applied to examine large data spaces (e.g. modern astronomical databases, or analysing the results of particle physics experiments). ‘Sophisticated’ here means that they require substantial effort to learn and understand. In contrast, in business applications one might find relatively simple and familiar methods (e.g. regression or cluster analysis) being used a huge number of times on different sets of customers or variables. For example, in building scorecards for predicting creditworthiness, large numbers of possible ways of segmenting the population are explored, along with large numbers of possible sets of predictor variables to combine when constructing a logistic regression model (say) in each segment. Of course, these descriptions of what goes on in different disciplines are generalisations, and one can readily

find exceptions. Note also that the terms ‘large’ and ‘a great many’ in all of this are relative: progress in computers has meant that a few years ago a ‘large’ data set might have contained just a thousand data points, whereas nowadays it might contain many millions or even billions.

It is useful, particularly in the context of data mining applications in forecasting, but also more generally, to distinguish between two different kinds of data mining exercises. I call these ‘model building’ and ‘anomaly detection’. Model building is an exercise with which all statisticians are familiar. The aim of modelling is to reduce the data set to a simpler description, which can then illuminate mechanisms or relationships, or can be used for exercises such as prediction or decision making. Time series models for forecasting are a familiar kind of model, but others include market segmentation for characterising likely future behaviour of customers, linear and generalised linear models for predicting outcomes, and so on. In contrast, in anomaly detection, the aim is to look for something unusual: the sudden departure from the norm, the extreme observation, the change in behaviour, etc.

When building models, one can often work with a sample. For example, basic laws of probability tell us that we may well be able to construct a model of sufficient accuracy using a (properly taken) sample of just 1000 data points, in place of the billion in the entire data set. Entire disciplines — survey sampling is an illustration — are built on this truth. In fact, however, there are subtleties. The size of the required sample will depend on both the accuracy one wants to attain and the complexity of the model one wants to build. To characterise a data distribution merely by its mean and variance requires a relatively small sample, but to describe also its skewness and kurtosis, along with other aspects, will require a larger sample.

As we push this notion further, so we reach the stage of trying to model very small features of the ‘true distribution’ from which the data arose. That is, we enter the realm of anomaly detection. Typically in data mining, however, this is approached from the other direction. Instead of trying to build a global model which describes the entirety of the underlying distribution, we focus on the data itself, and seek to detect unusual data points, groups of points, relationships between points, high frequency counts,

etc. Having detected such unusual configurations, we can then ask ourselves the inferential question: whether they could easily have arisen by chance.

At the very extreme, when we pose the anomaly question about individual data points — is this data point unusual? — we are forced to examine each and every data point. Sampling is of no potential use here (though sampling may be helpful in constructing a model with which each individual data point is compared — in outlier detection, for example). An example of such a situation would be detecting fraudulent credit card transactions, where sampling is clearly likely to be of little help. One simply has to examine each transaction.

Increasing numbers of large and very large data sets, along with the development of very fast computers facilitating the rapid exploration of data sets in many different ways, hold out immense potential for extracting meaning from data, and in particular for improved forecasts and predictions. However, such potential power does not come without its risks. I have already mentioned chaotic and stochastic limitations, but one must always be alert for other, rather more mundane potential problems. The next section describes some such.

4. Problems

The combination of large data sets and observational data mean that data mining exercises are often at risk of drawing misleading conclusions. In this section I describe just four of these dangers. These problems are certainly not things I alone have detected. Indeed, within the statistics community, they are problems which are well understood. However, the central philosophy of data mining — throw sufficient computer power at a large enough data set and interesting things will be revealed — has meant that they have often been overlooked in data mining exercises. Unfortunately, the solution is to temper the enthusiasm, and to recognise that rather more complex models are necessary. Statisticians generally do not build complicated models simply for fun, but for good reasons.

Problem 1. Selectivity bias.

I noted above that most data mining activities are carried out on observational data. By this I

mean that the researcher had no control over what treatments, exposures, or conditions the objects being studied were subjected to. This is in contrast to experimental studies, where such control is exercised. The risks associated with observational studies are well known. Primarily, because of the lack of control, and in particular the lack of opportunity for random assignment to different ‘treatment’ groups, there is a risk that observed differences between groups of objects may be due to unrecognised factors. For example, in a study aimed at identifying the distribution of different kinds of astronomical objects, dimmer objects are less likely to be detected. Since objects which are further away are likely to be dimmer, there is a relationship between proximity and probability of being detected. Then, because of the finite speed of light, we are observing further objects at an earlier time of their existence, so we obtain a time-distorted picture of the population of star and galaxy types across the universe. Things are then further complicated by interstellar and intergalactic gas and dust clouds, which attenuate radiation. A densely populated region of space may appear sparsely populated merely because the light is not getting through to us. Of course, all of these phenomena are well-known to astronomers, and appropriate adjustments are made in astronomical studies, but things are more difficult in situations where the data selection mechanisms are not so well understood, or, even worse, the possibility of such mechanisms is ignored.

A familiar case arises in the retail finance sector, where credit scores are used as the basis on which to make decisions about selling financial products (Hand, 2001a; Hand & Henley, 1997; Rosenberg & Gleit, 1994; Thomas, 2000). In particular, the aim is often to forecast whether an applicant for a product (e.g. a loan, a credit card, a mortgage, car finance, etc.) is likely to default on repayments within two years. A variety of different types of credit score have been developed, but typically they will include information on past behaviour with financial products (e.g. default on previous loans, slowness in making repayments, nature of credit products used in the past, fraction of credit limit reached, etc.), as well as other permitted characteristics which have been found to be predictive of the probability of defaulting (for example, whether or not a homeowner,

time with current employer, etc., but not including gender, which is prohibited by law). The last thirty years has seen considerable development of such models, and some highly sophisticated approaches have been developed. Although a wide variety of statistical and machine learning approaches have been investigated, including, for example, neural networks, random forests, support vector machines, and so on, by far the most popular type of model is a logistic regression tree, or ‘segmented scorecard’, as it is called in the industry. Such a model partitions the population of potential customers into segments (e.g., one might have three segments: those who have previously defaulted, those who have not and who have many existing lines of credit, and those who have not and have few existing lines of credit), and then builds distinct logistic regression models within each segment to predict default, using retrospective data.

Now let us look at this process from the perspectives of the data used to construct the model and the aims of the exercise. To do so, let us step right back to the beginning of the process. Our aim is to make the best predictions we can of the default risk of anyone who applies for the financial product — a loan, say.

In order to construct our predictive model, we need data describing previous customers, some of whom will have defaulted and some not. To obtain such data we started by soliciting applications for the loan. This might have been by direct mailing, via the internet, or by some other means. Amongst those who responded to the solicitation, many will have failed to reply. Amongst those who did reply, we will have used some earlier scorecard (or maybe even subjective judgement) to decide to whom to offer the loan. Amongst those who were offered the loan, only some would have taken up the offer (the others may have found the terms unattractive, or have changed their mind, etc.). And amongst those who did take the loan, some would unfortunately turn out to be bad risks and will have defaulted.

These various selection processes will have finally produced a population of customers who took the loan, some of whom defaulted and some of whom did not. This gives us a population which we can use to build a model to predict the likely outcome of a new customer, with known characteristics. However, this final population has undergone many selection

steps, and might be quite unlike the population of people who apply for a loan. In particular, just to take one aspect by way of illustration, the population whose outcome we have actually observed consists solely of people whom (we originally thought) would be good risks. Because of this, assuming that our initial suspicions were reasonably well-founded, the population whose outcome we observe is likely to be significantly less risky than the overall population of applicants: it will contain a lower proportion of people likely to default.

Of course, the consumer credit industry is well aware of this problem, and considerable effort has been made to overcome it. This effort goes under the name of ‘reject inference’ (Hand, 2001b; Hand & Henley, 1993), based on the counterfactual notion that one would like to determine the outcome class of those applicants whom one previously rejected for a loan if one had in fact offered them one, so that these outcomes could also be used when constructing the model.

The fact is, however, that the basic problem, as presented above, is an insuperable one: unbiased models cannot be constructed unless additional information is introduced. This extra information might come in various forms, including data about the earlier decision process or assumptions about underlying distributions, or from extra data. In fact, this example is an interesting one because, in contrast to many other data mining situations where such population selectivity arises, the problem has been recognised and has a known solution, at least in principle. The ideal solution is to change it from an observational to an experimental study. This can be done by accepting a random sample of applicants whom, one believes, are likely to default, in addition to accepting those that one believes are less likely to do so, so that a scorecard free from the distortions of sample selectivity can be constructed. This does require an enlightened understanding of the principles of experimentation, because it necessarily means selling the financial product to some people who would be regarded as a high risk. In my experience, most banks are uneasy about this. I regard this as a manifestation of a short term perspective: more overall profit can be made by sacrificing some short term gain in the interests of learning more about how customers behave.

Of course, there are exceptions. In particular, Capital One is renowned for its constant experimentation to discover the best products to provide for its customers. It is reported to carry out some 60,000 experiments a year. This immediately produces a very large data set (without even considering the individual responses of the customers within each arm of each experiment). To extract useful information from such a mass of studies, data mining tools are needed. However, at least as far as this paper is concerned, the key aspect of this experimentation is that it includes notions of a willingness to assign some predicted bad customers to the ‘good’ arm. Recognition of this principle has led Capital One to phenomenal success.

This example shows population distortion arising as a result of some prior data selection process, but it can also arise in many other ways. One very popular approach for handling incomplete data is simply to discard any incomplete records. As all statisticians know, this can be a dangerous strategy, since it risks leading to an analysis sample which has a distribution different from that of the complete population. Statisticians have developed a deep understanding of missing data, missing data mechanisms, when valid inferences can be made, and how to adjust for missing data, but data miners very often ignore it. The reject inference problem can, of course, be seen as a missing data problem, since there the higher risk applicants are disproportionately more likely to have been excluded from the sample with known outcomes which is available for analysis.

Problems of selectivity bias are not new. The potential for adversely impacting small data sets is just as great as with large data sets, but perhaps the dangers are less obvious with large data sets. I certainly think that data miners have been slow to address the issue. Moreover, if one is seeking small effects amongst large numbers of data points, the potential for these small effects to be caused by unrecognised influences is considerable.

Problem 2. Out of date data.

For sound pedagogical reasons, much statistics is taught from the batch mode perspective. That is, the analyst (a student, say) is given a set of data (complete, with no missing values, and assumed to be without errors, of course), and is requested to fit a model to it. However, the truth is that many analyses are

really conducted within a latent context of a stream of data, of problems, and of questions. In business, for example, I conjecture that almost all analyses are of this kind (and yet basic business statistics texts do not emphasise it), since businesses generally aim to continue into the future. Perhaps because much of the economic impetus for data mining has come from business needs, there has been considerable interest in what has come to be called *streaming data* in the data mining community. Such streaming applications are closely tied to forecasting — in most cases, businesses will want to use the information they acquire from an analysis to guide their future decision making. The point is that elaborate and sophisticated models do exist for coping with evolving data and problems (dynamic linear models, for example), but these are seldom applied in day-to-day data mining applications.

To illustrate, I again turn to the retail financial services sector, and credit scorecards.

To build a scorecard, we need both the potentially predictive characteristics (described above) and the outcome (e.g. default or not) of a sample of customers. Clearly, these are customers who were signed up some time ago, since we have had to wait until they have had the opportunity to default. Once again taking a loan as an example, in principle this means waiting until the end of the loan period. Suppose, for illustration, that the loan term is two years. Then, to be certain that someone will not default, we must wait until two years after they took out the loan to determine that their true class is ‘good’. If they do default before the two years are up, then we immediately determine that they are ‘bad’. However, we cannot choose a time less than two years and look at their status, or at least not without more elaborate analysis, or we would risk selectivity problems of the kind described above. This means that the predictor data on customers from whom we build the model relates to a population of customers which is at least two years out of date. This is not a problem if the system is stationary: if the distributions do not change over time (that is, if there is no ‘population drift’), but it can be a serious problem if they do. This point is particularly relevant at the moment, because we have recently had a long period of relatively benign conditions for consumer credit, which was suddenly brought to a crashing end with the ‘credit crunch’. Models built in the benign period may not be relevant

to the present, as is demonstrated by the dramatic increase in house repossessions and negative equity in recent months.

One expects the performance of predictive models to degrade over time as populations change. However, the above means that the models should be expected to have degraded before we start: they are (in the example above) already two years out of date before they are even used.

Various approaches have been explored for tackling this problem. They include:

- (i) survival analysis, where one truncates the observation period to less than two years, as suggested above, but makes explicit allowance for the fact that some of the customers may go bad after the observation date. Of course, there is a limit to how short an observation period one can take. At the least, since ‘default’ is often defined in terms of three consecutive months of failure to pay an installment, one may have to wait three months, and this may not yield sufficient defaulters to build a reliable model.
- (ii) dynamic logistic regression models.
- (iii) More elaborate models based on the hypothesis that there are characteristic types of customers, some more likely to default than others, and that this trait type is stable for a given customer. This allows the models to be split into two parts, a part relating the customer’s demographic and circumstantial characteristics to their type, and a part relating the type to the default risk. The first part is a very short term model. This part can be designed using old data, but when used it will be based on customer data which is only a few months old, so that it is very up to date. The second part is invariant over time: it can be based on old data, and will still be a valid link between the customer type identified in the first part, and the default risk.

However, despite the existence of these more elaborate models, the sector relies on relatively simple non-dynamic models, monitoring their performance until the degradation seems sufficient for them to need to be rebuilt.

In fact, this exposition merely scratches the surface of the difficulties. The aim of building predictive models in this sector, and indeed in business

applications in general, is not to see how clever we are at predicting the future, but is often to take some action in the present. Indeed, and in particular, it is to make some intervention. If, for example, we predict that someone is likely to fall into arrears with repayments, we might well contact them and arrange a revised repayment schedule. This very intervention will change the nature of the data on which the prediction is based, and so will invalidate the model. We have a reactive situation: what we do affects how the customer behaves.

In other domains yet further complications arise. Economic data are subject to revision as time progresses, because, for example, raw data takes time to come in, and more comes in as time passes. This means that the current estimates of things such as inflation, GDP, etc. are likely to be improved upon as they age. As a consequence, rather than the conventional approach of weighting the most recent data most heavily in forecasts, it can be better to weight older data more heavily.

Data miners (or at least those in business practice, rather than those who present ideas at academic data mining conferences) tend to stick to relatively simple methods — for example, repeatedly rebuilding scorecards, as noted above. The ability to do this is clearly another consequence of the computer revolution. One question I would like to raise is whether this is a good thing. As far as changing circumstances are concerned, it means that one’s models can adjust to the current situation, and avoids relying on possibly dubious assumptions which might be made by more advanced methods: survival analysis has to assume some distributional form to extrapolate into the future to decide whether someone is likely to default before the end of the loan term, and the bipartite model solution is making a fundamental assumption about the nature of customers.

Problem 3. Empirical rather than iconic models.

There are two distinct kinds of statistical models, which go under various names, but which here I will term *iconic* and *empirical* (Box & Hunter, 1965; Cox, 1990; Hand, 1985, 1994). Iconic models are mathematical representations (‘images’) of (necessarily simplifying) theories describing the phenomenon in question. Thus we might have a physical theory which tells us that objects will

accelerate as they fall towards the earth, and we might fit such a model to a set of data. Conversely, empirical models are based purely on finding convenient or useful summaries of a data set. Many regression models are of this kind – in a particular context there may be no theory saying that a mean response should be a linear combination of a set of predictor covariates, but a regression model may be used nevertheless. The relative balance of iconic and empirical models varies across disciplines and changes over time, and a model can start out as empirical and become iconic as understanding grows.

In general, I believe, iconic models should be expected to yield superior predictions to empirical models. That is provided, of course, that the models are ‘right’, in that they do represent important aspects of (or, perhaps, ‘good approximations to’) the way the system being modelled really behaves. The rationale behind this belief is the fact that models are generally composed of various components, so that one can think of these components as forming a set of basis functions by which to represent the system being measured. An iconic model (with the proviso above) is thus based on a good set of basis functions, which permits a reasonable approximation to the system, without extensive model searches and without the danger of adding superfluous basis functions simply because, by chance, they happen to fit well to the particular (finite) data set at hand. In contrast, an empirical model is either the result of a search over a much wider set of basis functions or is the result of a prior restriction to a particular set (e.g. a linear combination of predictor variables). In general, fitting a model using a smaller set of well-chosen basis functions leads to more accurate estimation — with, again, the proviso above, that this permits a reasonable estimate of the ‘truth’. Without this proviso there is a risk of bias, and hence inaccuracy of a different kind.

Now, the fact is that most predictive data mining models (in commercial applications, at least) are empirical. This is reasonable enough — in most situations there is little theory on which to base an iconic model. However, I think that there is a danger here. I call this the *cliff edge effect*: It describes the sudden dramatic deterioration in predictive model performance.

Empirical models are well-matched to the data at hand. They describe the retrospective set of

data available for constructing the model very well, and, if carefully built, allow accurate generalisation and prediction of new cases drawn from the same distribution. However, as we have already seen, in the credit arena, and I would conjecture in most other business applications, population drift (driven by changing economic circumstances, competitive environment, technological progress, etc.) means that the new cases are not drawn from the same distribution. Indeed, we have already seen that, in the case of credit scorecards, the forecasting model is some years out of date before its usage even commences. In such cases, I suspect that iconic models (again with the proviso above) will be more reliable, and less subject to the cliff edge effect.

Of course, constructing good theories on which to base one’s iconic model may not be easy. However, it is possible to step partly towards this ideal. Once again, to illustrate, I turn to the credit scoring domain.

As we have already seen, in this domain, the models are generally empirical. They collate a set of possible predictor variables, measure the associated outcomes (e.g. default or not), and then build the segmented scorecards or whatever, purely on the basis of an empirical analysis identifying associations in the data. The resulting model has the familiar regression form

$$r = f(x_1, \dots, x_p),$$

where r is the outcome, x_i are the predictor variables, and f is a function permitting the prediction of the outcome from the predictors. f or its parameters are estimated directly from the retrospective data, which includes observations of both x_i and r . Note that in such a model no restrictive assumptions are made about the relationships between the x_i . Typically logistic regression models or logistic regression trees are the chosen form for f in the retail credit industry, and extensive data mining work is used to construct them, including trying different segmentations, different transformations of predictor variables, and different sets of predictor variables.

However, there is an alternative to this. In particular, one can conceptualise ‘creditworthiness’ as a latent variable, a characteristic of the customer. This will be intrinsically unobservable, but will be influenced by the primary characteristics of the customer, and will in turn influence various behavioural characteristics. For example, denoting this latent characteristic by q (for

‘quality’ — see Hand and Crowder (2005)), variables such as age, socio-economic group of parents, education level of parents, and so on, might be regarded as primary characteristics — they influence, but are not influenced by, the creditworthiness of the individual being rated. In contrast, examples of behavioural characteristics would be arrears history and current account history. We might reasonably regard these as potentially influenced by creditworthiness. Certainly they are of a qualitatively distinct kind from the primary characteristics. Moreover, we might reasonably assume that these behavioural characteristics are conditionally independent, given the q value. That is, any relationships between the behavioural characteristics are induced by their mutual relationship with q . This thus yields a more elaborate *multiple-indicator-multiple-cause* model, with q being the unobservable but latent variable in the middle, which can be estimated (Hand & Crowder, 2005).

The point of this is that it is a step away from the purely empirical models traditionally used in data mining, towards an iconic model form. Here the theory is very weak — merely saying that certain aspects of an individual influence the latent q variable, that others are influenced by it, and that these latter are conditionally independent given q . However, it is a first step.

Problem 4. Measuring performance.

The final problem I want to mention is that of measuring the performance of predictive models, and the closely related issue of the criterion used to choose between models. The starting point is the self-evident truism that different measures lead to different models being chosen. Models for predicting a binary prognosis of hospital patients based on optimising the misclassification rate are likely to be rather different from models chosen on the basis of likelihood. There is nothing deep in this: ranking people by weight is likely to yield an order different from ranking them by height, even though one might expect the two rank orders to be correlated.

Since different performance criteria are likely to yield different orders of merit, it is clearly important to choose a criterion which closely matches the objectives of the analysis, and yet this is often not practised. All too often, too little thought is given to

the ultimate objectives of an analysis, and a criterion is adopted by convention.

Sometimes there are sensible practical reasons for avoiding choosing the most appropriate criterion, though the risks are not always appreciated. For example, in predictive classification problems (like the binary prognosis forecasting problem mentioned above), the misclassification rate is often chosen as a performance measure (indeed, in comparative evaluations of such methods by the data mining, machine learning, and statistics communities, by far the most common criterion is the misclassification rate, see Jamain, 2004). However, the misclassification rate is rarely chosen as the criterion to be optimised when determining the model. This is because, being discrete, it is difficult to optimise. Instead, more typically, a measure such as the likelihood is chosen.

One can take this further. Even though the misclassification rate is a common evaluation measure for binary prognosis problems, it is rarely an *appropriate* measure. More commonly, different kinds of misclassifications carry different relative degrees of severity, and this should be taken into account when choosing a criterion.

Taking this further still, practical experience shows that determining these relative degrees of severity is difficult. This has led to a variety of measures such as the Gini coefficient (or, equivalently, the area under the ROC curve), or partial areas under the ROC curve (e.g. McClish, 1989), which aggregate different relative degrees of severity. These are rather unsatisfactory measures, either because they make latent assumptions about the relative severity, or because they make these explicit, and hence introduce subjectivity. Likelihood, on the other hand, is universal (all researchers, applying the same model to the same data, will obtain the same likelihood). Of course, if one is prepared to believe that the family of models being contemplated includes the ‘true’ model, then any measure of discrepancy between the true model and the fitted model can be defended — and likelihood has many attractive properties. On the other hand, if the assumption is difficult to justify (as must surely be the case in all empirical modelling) then it seems less acceptable. Hand and Vinciotti (2003) explored this issue.

5. Conclusion

Forecasting is fundamentally an inferential problem. That is, it is not simply a question of summarising data, but is rather a question of generalising from the available data to new data — and in particular to new situations which are likely to arise in the future. In contrast, the early development of data mining by the computer science community put emphasis on the analysis of the data set to hand (e.g. the discovery of ‘frequent itemsets’ in large transaction databases). It is only relatively recently that the inferential nature of many of the problems addressed by data miners has been properly recognised. Inference is a much tougher problem than summarising. It requires careful thought about how the available data arose, so that one can be sure that one has a properly representative data set, permitting the powerful tools of probability and statistics to be properly applied. I suspect that, all too often, certainly in the past and to a large extent in the present, such issues have been overlooked by data miners. It is, perhaps, a tribute to the power and potential of data mining that, despite these dangers, the discipline has gained in importance and reputation.

With this as a background, I believe that data mining is changing. The central importance of inference to many of its concerns is being recognised. Moreover, although many data analyses are based on retrospective observational data originally collected for some other purpose, increasingly we are seeing data mining ideas being applied in experimental settings. This holds the promise of very exciting developments in the future.

Data mining, in commercial practice at least, is often characterised by the extensive fitting of relatively simple models. Moreover, these are almost universally empirical. Empirical models have the strength that they might include a powerful predictor which an ‘expert’ would not have recognised as relevant — they can include things we would never have thought of — to increase the predictive power. However, this is not without its risks. In particular, empirical relationships are susceptible to the cliff edge effect, and the predictive performance may degrade dramatically if relationships alter as the circumstances surrounding new data change. Also, empirical models, while they

might lead to effective prediction and forecasting, do not lead to an enhanced understanding of the underlying truths.

Acknowledgements

The author’s work on this paper was partially supported by a Royal Society Wolfson Research Merit Award.

References

- Box, G. E. P., & Hunter, W. (1965). The experimental study of physical mechanisms. *Technometrics*, 7, 57–71.
- Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science*, 5, 169–174.
- Hand, D. J. (1985). *Artificial intelligence and psychiatry*. Cambridge: Cambridge University Press.
- Hand, D. J. (1994). Deconstructing statistical questions (with discussion). *Journal of the Royal Statistical Society, Series A*, 157, 317–356.
- Hand, D. J. (2001a). Modelling consumer credit risk. *IMA Journal of Management Mathematics*, 12, 139–155.
- Hand, D. J. (2001b). Reject inference in credit operations. In E. Mays (Ed.), *Handbook of credit scoring* (pp. 225–240). Chicago: Glenlake Publishing.
- Hand, D. J., Blunt, G., Kelly, M. G., & Adams, N. M. (2000). Data mining for fun and profit. *Statistical Science*, 15, 111–131.
- Hand, D. J., & Crowder, M. J. (2005). Measuring customer quality in retail banking. *Statistical Modelling*, 5, 145–158.
- Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry*, 5, 45–55.
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, 160, 523–541.
- Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, Mass: MIT Press.
- Hand, D. J., & Vinciotti, V. (2003). Local versus global models for classification problems: fitting models where it matters. *American Statistician*, 57, 124–131.
- Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, 2, 167–191.
- Jamain, A. (2004). A meta-analysis of classification methods. *Ph.D. Thesis*, Department of Mathematics, Imperial College London.
- McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, 9, 190–195.
- Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research*, 42, 589–613.
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16, 149–172.