An Introduction to Data Mining

Christopher R. Stephens[1,2,3,4] and R. Sukumar[5]

[1] Adaptive Technologies Inc.,
N. 67th Street, Peoria, AZ
[2] Instituto de Ciencias Nucleares, UNAM
Circuito Exterior, A. Postal 70-543, México D.F. 04510
[3] Dublin Institute for Advanced Studies
10 Burlington Road, Dublin 4, Ireland
[4] Department of Computer Science
University of Essex, Wivenhoe CO4 3SQ
[5] IPSOS

1 Introduction

An exponential explosion in data has occurred in the last 20-30 years, in basically all areas of business, due to the widespread availability of powerful computers for data storage and analysis. This has ensured an importance for the field of data mining, or as it is sometimes called, and more appropriately so - knowledge discovery - that can only increase in the future.

It should be emphasised from the outset, that data mining is not a tool or technique but, rather, an entire field of study. In fact, one could argue that data mining encompasses elements from the vast majority of other topics covered in this book - ranging from assuring data quality (see [1–5]) and initial basic data analysis (articles [6, 7]), to the use of sophisticated analysis tools (as, for example, in [8, 12, 13]). We make no claims as to proffering a comprehensive overview of the field,[1] but, instead, will try to paint a complementary picture to that which appears in most standard treatises.

"Googling" the words data mining gives a large number of competing, though more often than not similar, definitions. We consider data mining to be the "exploration and analysis of data in order to discover patterns, correlations and other regularities". To discover and analyse such patterns and regularities, one uses data mining models as representative templates to model them. The range of models that can potentially be applied is quite vast; basically encompassing any statistical model, pattern recognition or search technique - ranging from the simple, such as a univariate linear regression, to the very sophisticated, such as a neural network with weights found using a genetic algorithm.

While accepting data mining as the search for patterns in data, here, we will frame our discussion using a novel conceptual viewpoint that characterises

[1] In the appendix we present a brief bibliography of some useful data mining resources, including books [14–20] and on-line resources [21, 23, 22, 24], as well as an academic journal [25] dedicated to data mining, in order to allow the reader to delve more into what is already a very large literature.

this exploration as the search for "*Predictability*", a quantity which is a measure of the degree of "reproducibility" of the patterns and regularities within data. Thus, by observing patterns and regularities in one data set, we wish to predict what patterns and regularities will exist in another, statistically similar data set. Predictability varies as a function of a problem's "features" and "feature values", i.e., the predictor variables in a problem and their values[2]. For instance, one might find that income is a weak predictor of car ownership but a strong predictor of luxury car ownership.

Predictability can be very multi-faceted, reflecting the multi-objective nature of real-world business problems. In this context, there is no "magic bullet" data mining model. Rather, each model has its own characteristic advantages and disadvantages; situations where it works well, and those where it doesn't; objectives that it achieves well and others that it doesn't. Therefore, data mining really requires a "multi-perspective" approach, whereby various data mining models are used, in order to satisfy such diverse problem requirements as: precision, accuracy, transparency, understandability, implementability, simplicity, etc. Using more than one data mining model offers many benefits. For instance, in order to reduce variance in the presence of noisy data, one could search for a statistical model that depends on a set of parameters, and that describes the overall data, by dividing the latter into multiple data sets. One then finds parameter values for the overall model by taking a (weighted) average of the parameter values found on the individual data sets. Of course, the use of multiple models may be more sophisticated than this. One may, for example, use a neural network to generate predictions of which customers are more susceptible to churn. However, neural networks are notoriously difficult to interpret, so one may require a further analysis with a more simply interpretable model, such as a logistic regression, in order to offer a simply interpretable logic to attach to the predictions of the neural net.

Although our emphasis is on data mining models, we will spend some time discussing other aspects of data mining, such as data acquisition. Although this is discussed elsewhere in this volume, we want to emphasise that the two are inextricably linked. Just as data mining is the search for Predictability, so the data acquisition process determines precisely just how much Predictability is present. If there is little or no Predictability in the data, then no data mining technique can overcome that.

We will also try to inform our discussion by referring to various insights, rules-of-thumb, etc. that we have gleaned from various real-world data mining projects in which we have been involved.

[2] There are several different terms used for the input variables in data mining, depending on the principle domain of expertise of the practitioner. In this article, we will use the term "feature" to describe an input variable, i.e., one of the variables to be mined. Other equivalent terms are attribute, or characteristic.

2 What Is Data Mining?

As mentioned, our definition of data mining is: the exploration and analysis of data in order to discover patterns, correlations and other regularities - as characterised by the search for Predictability. This can be approached in two principle ways: First, one may posit a particular model to describe a pattern, then test it and validate it on the data. This is the "top down" approach. Alternatively, one may discover a pattern without any a priori model in mind, by letting the data "speak for itself".

The advent of cheap and powerful computing resources means that the data analysed is almost inevitably electronic and, more often than not, comes in large quantities. The ultimate objective of the data mining process might be economic - as probably you have in mind - or scientific. Wherever large amounts of data have to be analysed, it is an area of crucial strategic importance. It is important to emphasise though that, unlike the rest of the more technical contributions to this book, it is not a technique or tool; rather, it is an entire field of study which can be applied to many different areas of business.

2.1 The Goals of Data Mining - Profiling and Prediction

Data mining, as the discovery and analysis of patterns and regularities, can be related to two broad classes of tasks: Profiling and predicting, which, in their turn, are related to the ideas of pattern idnetification - in the case of predicting - and pattern description - in the case of profiling. Typical prediction tasks are centred on the key questions of: Who? What? Where? and When? - such as: Who is most likely to be delinquent in their future payments? What is the predicted average cost of acquiring new customers? Where should a new store be built to optimise ROI? When is someone likely to stop being a customer? Profiling on the other hand is more concerned with what a particular group looks like, and is very much associated with the Why? behind the above questions. Generally, we can envision profiling those who are members of a given class. For example, those customers who provide the most ROI, or the most loyal customers, or those who are most likely to buy an upgrade on a product, etc.

In answering the questions: Who? What? Where? and When?, almost inevitably, there is no degree of certainty about the answers. The world is inherently "noisy", and this is especially true in an area such as marketing, where a complex world driven by human behaviour is being modelled. Under these circumstances, it is better to not just provide an answer to these questions, but rather a *probability* for a given answer. So, for example, in predicting who will be delinquent in their future payments, it is better to determine the probability that a certain customer will be delinquent. Similarly, for predicting the cost of acquiring new customers, more relevant than just the average is the probability distribution. Perhaps there are potential customers that cost very little to acquire, and others that cost a lot, leading to the same average as in the case where every new customer costs about the same. In other words, the distribution of costs is potentially very important. In this case, what is of interest is $P(C|X)$, the

probability for C given $\mathbf{X}$, where C is the output variable and $\mathbf{X}$ represents the set of input variables (features). For a given sample, $P(C|X) = n(C, \mathbf{X})/n(\mathbf{X})$, where $n(C, \mathbf{X})$ is the number in the sample that are in the class C *and* are associated with the input variable values corresponding to $\mathbf{X}$, while $n(\mathbf{X})$ is the number in the sample associated with $\mathbf{X}$, irrespective of whether they are in the class C or not.

Profiles are very often associated with determining the key drivers associated with membership of a certain class, such as the top decile for ROI, or bottom decile for churn. The key drivers may be single features, or feature values, or combinations of them. The richer a profile, the more informative it is likely to be. Due to the multi-objective nature of business, it is important to emphasise that prediction and profiling are not orthogonal objectives, but rather must be pursued simultaneously. It is not sufficient to simply have a list of those customers most likely to purchase an upgrade of a product. One must also understand what these customers look like, how to approach them, what is actionable in the light of this list, etc. - especially if one wants to convince upper management.

As well as being associated with distinct, but complementary and equally important, objectives, predicting and profiling often use quite different models. As profiles are what determine the logic behind a certain observation or prediction, certain modelling techniques are less suitable than others. As mentioned, neural networks, for instance, can be a useful non-parametric statistical estimator. However, their lack of transparency and interpretability makes them less suitable for profiling.

3 What Data? - Plan to Predict

It is important to realise that data mining is only as good as the data you have. If you start with dross, no matter how hard you work, it will remain dross. However, if one starts with some reasonable raw material, then it will be more feasible to search for those data mining nuggets of gold that one wishes to discover. Striking gold in data mining means finding *Predictability*. As mentioned, this can be thought of as a quantitative measure of the relatedness or correlation between a desired output variable - a prediction or a profile - and a set of input variables - feature values or features. We shall discuss Predictability at greater length in the following section. Here, we concentrate for a moment on how choice of data and the data collection process should be done. This should always be with a careful eye on the post collection data mining process so as to optimise the chances of finding Predictability. We call this at Adaptive Technologies - *"Planning to Predict"*.

Frequently, the data that one is required to mine is a preexisting data set. In this case, its content and parameters were often set without any reference to any future data mining analysis. This can cause grave difficulties, as it may be that the data is simply not predictive with respect to what one wishes to predict. In other words the data is dross. However, how does one distinguish dross? Might it not be that one is looking for gold in the wrong place, or that the tools

one is using are inadequate? There are however, rules of thumb that should be implemented in order to reduce the chances of ending up with data dross. There are several aspects that can affect the a priori Predictability inherent in a dataset. We will now consider some of them.

3.1 Domain Specific Knowledge

In the Plan to Predict philosophy, one tries to incorporate domain specific knowledge into the process of what data to collect. For instance, how does management articulate the particular business goals of the company that are to be met by collection and analysis of data? This, after all, can be a costly process. What do people on the ground think are important variables driving customer loyalty, ROI, churn, etc.? What do your important sales people think are important characteristics of loyal customers? They should have a better intuitive feeling than someone from outside the company. Design for data collection should have as a focus taking key insights from company employees that can be translated into parts of data acquisition instruments, such as questionnaires.

3.2 Spatial and/or temporal inhomogeneity

Another important element that degrades the Predictability present in a data set is spatial or temporal inhomogeneity in the statistical distribution that underpins the data. In other words, one should compare statistically similar samples when comparing data across different time periods, or across different geographical (or, by implication, socio-demographical) areas. A related important point for predicting the future is: how much of the past is predictive of what will happen? This is very much related to the problem of whether or not a non-stationary probability distribution underlies the collected data.

These points can also be considered in a more formal context, using two key concepts familiar from statistical analysis - bias and variance. These can appear both in the sample from which the data were taken, as well as in the content of any data collection instrument. We have frequently seen strong bias associated with the statistical samples used, such as using statistically very different samples in different time periods, or combining data from very different types of collection point. For instance, using an in-store survey for customer data and combining it with mall-based surveys for non-customer data, only to find that the socio-demographic characteristics of the two different samples are completely different, thus introducing two different samples with strong but distinct biases. This requirement of statistical homogeneity also applies to the data set that one is going to use to train one or more models and the data set on which one wishes to predict. It is no good formulating a data mining model, however good it is, on one data set, when the data that one wishes to predict is statistically very different.

3.3 Important Spatial and Temporal Scales

Another crucial element in the Plan to Predict process is associated with determining the important spatial and temporal scales of a problem. For example, in trying to determine who to target in a mass direct mailing campaign, predicting who to target using zip code level socio- and psycho-demographic data might be adequate. However, if one is considering a much more costly intervention, such as a visit by a salesperson, then household level data would be much more predictive. Thus, the level of "granularity" of the data is a very important element in determining Planning to Predict. Similarly, for time scales - if we wish to predict sales with a time horizon of five years then looking at "real-time" point-of-sales data is not appropriate, rather the data should be "coarse grained", i.e., binned, over a time scale that is suitable for the time horizon for prediction.

3.4 Uncertainty and Predictability Barriers

Uncertainty is the flip side of Predictability. However, it is useful to distinguish between several different types of uncertainty: data errors; limited sample size; wrong choice of data variables; and changes in the underlying probability distribution. Data errors are almost self-explanatory. Similarly, the fact that small sample sizes lead to larger sampling errors, and therefore increased variance, is something readily understood. In principle, neither of these offers a fundamental barrier to how much Predictability may be obtained from the data. The next two however, play a much more subtle and important role. The goal is to establish a relation between a target - buying a product for instance - and a set of features that will correlate to that target. However, the space of potential features is infinite, hence, one must first make a reduced choice of variables from which to look for Predictability. Different choices of variables will be associated with different degrees of Predictability, there being for that set a Predictability "barrier" - a limit to uncertainty for a fixed target and feature set that cannot be overcome. Finally, uncertainty can arise because the underlying probability distribution has spatial or temporal dependencies as discussed above in section 3.2.

4 Elementary Data Mining

People have more experience of data mining than they imagine. In this section, we will show that some well known tools, seen in other chapters of the book, applied to problems of low dimensionality (few features), can be seen as exercises in elementary data mining.

4.1 A One-variable Example

Consider, for example, the (fictitious) data in Figure 1 relating the ROI for a given amount of spending on advertising in marketing campaigns across a

raft of different companies. Two companies, A and B, with this data in hand, wish to decide on whether or not a new \$5 million advertising campaign will generate enough ROI to make the investment worthwhile. The VPs of Marketing for the two companies are asked to predict what the likely ROI is. The VP of marketing of Company A (VPA) feels that there is a high degree of noise in the data and decides to trust in a simple linear regression of the form $y = a_1x + a_0$, finding a good fit ($R^2 = 0.83$) with the relation ROI = 4.0596Advertising − 5.1065. Meanwhile, the VP of Marketing of Company B (VPB) feels that a more sophisticated approach could work, and uses a non-linear regression of the form $y = a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, finding a "superior" fit ($R^2 = 0.90$) by fitting with the sixth order polynomial ROI = $-0.0007(\text{Advertising})^6 + 0.0409(\text{Advertising})^5 - 0.866(\text{Advertising})^4 + 8.7596(\text{Advertising})^3 - 42.836(\text{Advertising})^2 + 92.591\text{Advertising} - 55.966$. In both cases, a simple least squares regression technique was used to determine the coefficients a_i. From the analysis, the VPA predicts the ROI on a \$5 million campaign to be \$15.2 million, i.e., the value of ROI from ROI = 4.0596Advertising − 5.1065 when Advertising = \$5 million, while the VPB estimates it to be \$6.1 million using the sixth order polynomial. The result is that Company A goes

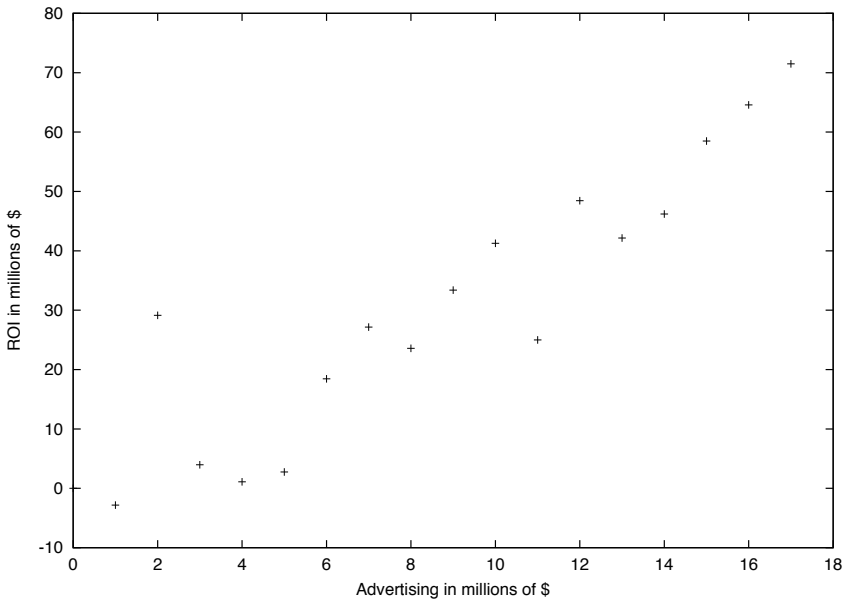

Figure 1. Example of simple one-dimensional data mining: determination of the relationship between advertising spending and ROI.

ahead with the advertising campaign and makes \$16.4 million while Company B decides against it.

This sort of problem represents an elementary form of data mining, the problem here being to predict an output variable, ROI, based on only one input, the predictor variable or feature - amount spent on advertising. Both input and output are represented by continuous metric[3] variables. In this case, a large natural class of data mining models are those defined by regression techniques. From this large class, VPA decided that there was not enough data to warrant more than a simple model. He chose a univariate linear regression, a model with few parameters (slope and intercept) to be adjusted. His rationale for this was that such a model offered a transparent, easily understandable relation between the two variables. VPB on the other hand, had access to more expensive statistical software and decided to explore a larger class of models, but still keeping within the regression class. After trying many different models, a non-linear regression within the class of sixth order polynomials was chosen. Comparing with a linear regression, he decided that the more sophisticated model offered more predictive power, even if it was not as easily interpretable. So, both VPs are being "data miners", using different criteria for choosing data mining models from within well known classes.

So, what can we learn from this simple example? Well, the underlying data was generated using $y = 3x + 2 + 20\xi$, where ξ is a random number chosen from the interval $-1/2$ to $1/2$. Thus, the underlying data was linear with a noise term. Of course, neither VP knew what was behind the data. In retrospect, VPA made the right choice of model, but was this luck or was VPA a better data miner? There were only 17 data points. For the linear model there are two free parameters ($a_0,\ a_1$) and, hence, only 15 of the data points are independent. However, the sixth order polynomial has seven free parameters. In this case, there are only 10 independent data points. This is, of course, a recipe for "overfitting". The more complicated model better fits the data, but only because it has keyed in on the noise inherent in the data. This could have been confirmed by increasing the number of data points. For instance, if the number of data points had been increased by a factor of 5, and both VPs had studied the data again using their respective regression models, then one would find that the statistical significance of the two models was almost identical. Further, one would find that the coefficients of the higher order terms in the sixth order polynomial of VPB were now much smaller, meaning that the curve of VPB and the straight line of VPA are much more similar when compared to the case when there are only 17 data points. Hence, after obtaining more data, VPB realizes that it was an error to use the model he used and that a linear regression would have been a better choice.

Now, consider the data seen in the left hand graph of Figure 2. Confronted with this data VPA, using the same logic as before, uses a linear regression, finding ROI $= 40.653\text{Advertising} - 82.042$ with an $R^2 = 0.78$. VPB, chastened by his previous experience, now also adopts a more conservative approach

[3] Metric here refers to the fact that the values are ordered with some associated notion of "bigger" and "smaller". Some common examples are: age, income, spending, price, etc.. Common examples of non-metric variables are ethnic origin, zip code, etc..

and uses a linear regression. Both VPs obtain the same predictions. Upon obtaining more data however, it becomes more apparent on examining the right hand side of Figure 2 that the relationship is not linear after all, but rather is non-linear, a good fit being possible with a quadratic expression of the form $\text{ROI} = 2.0422(\text{Advertising})^2 + 5.4626\text{Advertising} + 17.524$. This example shows

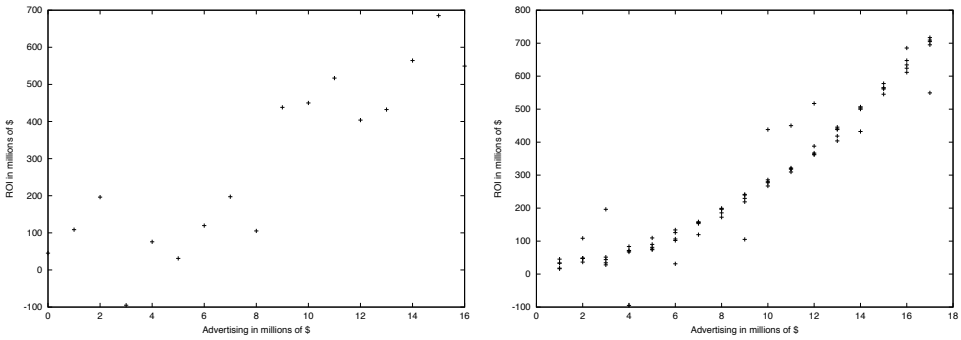

Figure 2. Example of simple one-dimensional data mining with non-linear relationship between output and input variables. Left hand figure is with small sample and right hand figure with larger sample.

the limits of a simple linear regression. Put simply, if there are non-linearities in the data then a linear regression might be quite inadequate. Regression techniques, passing to a larger class of models - polynomials, such as quadratics, cubics, etc. - can still be usefully used, but with the caveat that great care must be exercised not to over-fit, i.e., use a model with many parameters when the number of data points is small. Further, it may well be that a class such as polynomials in inadequate. If, for instance, there is seasonality in the data - something that is quite common in time series analysis - then a polynomial would not be appropriate. For example, considering the increase in toy sales in December. Thus, for predicting sales, a regression on a "customised" function that takes into account trends and seasonality might be necessary.

So, even at this very simplified level of one output and one input variable, one sees the myriad subtleties and complications that can occur in even relatively simple data mining tasks when trying to understand data and choose an appropriate data mining model. Although many aspects of data mining can be automated or semi-automated - such as a search through a class of models for a "best fit" - more often than not, there is no substitute for human intelligence and experience. It is important that an intuitive understanding of the data be obtained, rather than a superficial, or even erroneous, one gleaned from inappropriately applied overly-sophisticated and advanced data mining techniques.

4.2 A Two-variable Example

In the previous section, we saw the subtleties of data mining with only one feature, or input variable. Now we pass to the slightly more complicated case of two features. In Figure 3 we consider the case of two discrete metric features and an output variable $P(C|\mathbf{X})$ which represents the probability that the output variable takes the value C given values for the input features - feature 1 and feature 2 - represented by $\mathbf{X}$. In this case, our data mining VPs from section 4.1

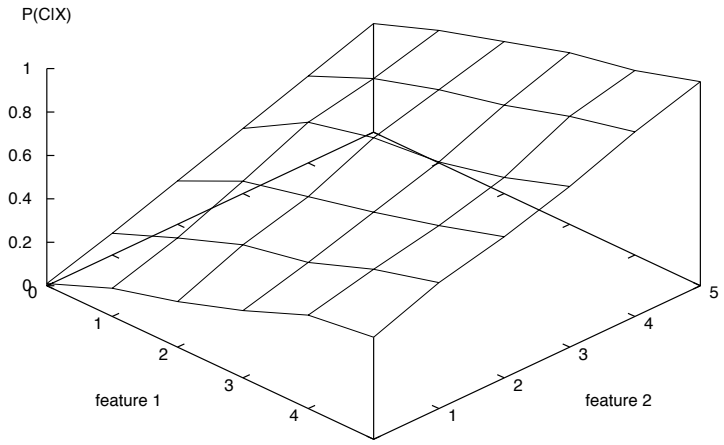

Figure 3. Example of simple two-dimensional data mining: determination of the relationship between output variable $P(C|\mathbf{X})$ and two discrete, metric features - a linear case.

would naturally use a multivariate linear regression, fitting a model of the form $P(C|\mathbf{X}) = a_{11}(\text{feature1}) + a_{21}(\text{feature2}) + a_0$. Here, the goal is to predict for a person with particular values of feature 1 and feature 2, what the corresponding value of $P(C|\mathbf{X})$ is.

Now imagine that the data was as in Figure 4. It is clear that a multivariate linear regression would not be suitable in this case. In fact, neither would any simple function be a particularly good approximation. The relation is multimodal, i.e., there are several maxima and minima, and, in fact, in the middle there is an isolated "mountain peak". The question is: how does one model this relation? Well, if there is sufficient data to establish, at least with some degree of confidence, the relation between the output variable and the input variables, then one can simply let the data "speak for itself"; i.e., the predicted value for

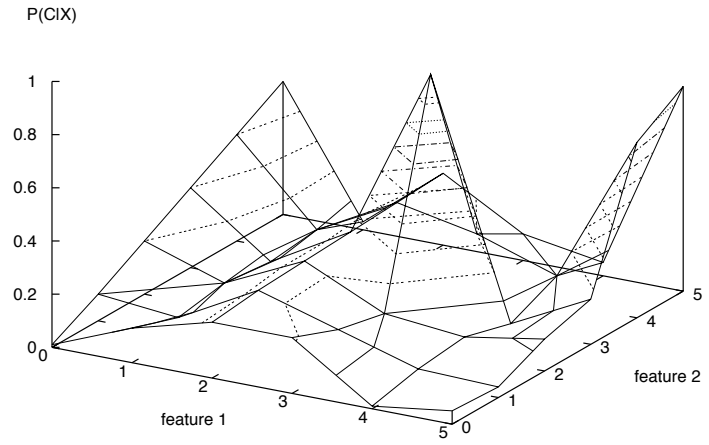

Figure 4. Example of simple two-dimensional data mining: determination of the relationship between output variable $P(C|\mathbf{X})$ and two discrete, metric features - a non-linear case.

the output given an input, would be simply that determined from the already established relationship. This means that one requires several observations for every single possible combination of feature values. This is feasible in the simple case where there are only 36 possible combinations, and say 200 observations. Imagine, for example, a two question survey with 6 responses per question and 200 respondents. The trouble is, real world data mining problems are not like this. On the contrary, there are usually at least tens, more often than not hundreds, and sometimes even thousands, of possible features. The corresponding number of possible feature value combinations is truly astronomical compared to the number of observations. This means that for the vast majority of combinations we have *no* observations! What can be done then? That is what we will confront in the rest of this article, and will constitute the difference between elementary data mining, as seen here and with which most people are familiar, and "real-world" data mining.

5 The Predictability Landscape

We have emphasised that data mining is intrinsically linked to the search for Predictability. Two important characteristics of this search are: i) What are the characteristics of the feature space being searched? and ii) What measures of Predictability should one use? - the overall goal being to determine which

features and feature values are predictive with respect to a given measure. If one thinks of this measure as a "height function" above the space of feature vectors, then the resultant space consists of a topography - the "Predictability Landscape".

5.1 The Curse of Dimensionality

As alluded to at the end of section 4.2, the most concise answer to the question about the characteristics of the search space is that it is enormous. Typical data mining problems are associated with hundreds or even thousands of features of many different types - socio-demographic, attitudinal, customer purchasing history, etc. - each with their characteristic feature values. The total number of potential feature value combinations is extremely large, even for relatively small problems. Put more quantitatively: For a set of N features (number of dimensions of the space), $\{x_i\}$, $i \in [1, N]$, each with feature values (number of points in a given dimension), x_{ij}, $j \in [1, a_i]$, x_{ij} being the jth value of feature i, a feature vector, $\mathbf{X} = x_{1i}, x_{2j}, \ldots, x_{Nk}$, is a set of N feature values, one for each feature. The total number of possible feature vectors is $n = \prod_i a_i$, where a_i is the number of possible feature values for feature i. For only 40 binary valued features, the number of possible feature vectors (2^{40}) is over one trillion - 10^{12}. There are two immediate consequences of this combinatoric explosion: Firstly, one requires an intelligent search heuristic for exploring the space of possible feature vectors; and, secondly, one needs to implement a "coarse graining", whereby the effective dimensionality of the search space is reduced. An intelligent search heuristic is required, as a random enumeration of all the possible feature value combinations would be inordinately expensive computationally. A very powerful computer evaluating the predictability of a million possible combinations per second would still require over 35 years to exhaustively check all possible combinations for only 50 binary features.

There are many available heuristics [26] for searching high-dimensional spaces, ranging from simple local search techniques, such as hill climbing, to more sophisticated ones, such as evolutionary algorithms [27, 28], TABU search [29], simulated annealing [30], etc. Each has its own characteristic advantages and disadvantages. Similarly, there are problem classes where one type of heuristic out performs another, and other classes where it under performs. The actual performance may also depend sensitively on various parameters that are input into the heuristic. From this point of view a practitioner who does not have ample experience with these techniques would be better advised to opt for a particularly transparent one, such as a stochastic hill climber. Essentially, the subtlety of using a search heuristic is that of striking an adequate balance between exploration and exploitation of the space of feature vectors. Too much exploration means that effort was wasted looking for new predictive features after the most predictive had already been found. On the other hand, too much exploitation means that the search spends too much time on a particular set of predictive features without realizing that there exist even more predictive feature combinations elsewhere in the space.

5.2 Coarse Graining

The second, and even more challenging, problem of having a high dimensional search space, is associated with making statistical inferences from the data. Thinking of the above example in the context of a survey with 40 binary questions - imagine having as goal trying to predict how much someone would spend, Y, in the next year on a certain product, based on their responses to the survey, $\mathbf{X} = \{x_1, ...x_{40}\}$. Thus, we are required to determine $Y(\mathbf{X})$. Now, one may posit a simple data mining model, such as a linear model $Y = \sum_{i=1}^{40} m_i x_i + c$. From the data one is then required only to determine the 40 gradient parameters, $\{m_i\}$ and the intercept, c. Sounds easy. However, first of all, one is positing a simple (only 41 parameters in a 40-dimensional space) but extremely biased model. In principle, one must determine to what extent this linear model fits the data better than any other possible model. Assuming that the relationship between Y and $\mathbf{X}$ were deterministic, i.e., only one unique value of Y for every $\mathbf{X}$, then to fix the form of the relationship and eliminate error completely, one would need to specify a value of Y for every possible $\mathbf{X}$, which would mean having over a trillion completed surveys! In reality, the situation is even worse, as the relations in data mining are never truly deterministic. There are always uncertainties. Clearly, a person's spending intrinsically depends on many more variables than the 40 encountered in the supposed survey instrument! This means that for a given $\mathbf{X}$ the corresponding Y values form a statistical distribution. Hence, one needs enough completed surveys not only to determine a possible value for Y for every $\mathbf{X}$, but also to have enough data to be able to estimate the probability distribution of Y values for a *given* $\mathbf{X}$. If one wishes to determine the distribution with absolute precision then an infinite number of surveys need to be completed for a given $\mathbf{X}$, i.e., each with the same answers.

Clearly, if the situation were as bad as we have naively painted here, then data mining would be impossible. The way to ameliorate these problems is by considering a "coarse graining". Coarse graining is a familiar concept to everyone, but not necessarily with that terminology. Put in a cruder but more familiar way, what is the optimal way to "bin" data, or group data together, or average over certain data fields or, in its most extreme form, to ignore some data altogether? Coarse graining enters into data mining problems at very many different levels. The most simple and elementary one is precisely that of binning data, for example, age. Keeping with the concrete example of a marketing survey: imagine that we have only 100 respondents, in which case one might expect less than 3 or so respondents of a given age in years[4]. This means that statistical inference based on such a small sample for a given age would be very difficult. However, would one expect to see significant differences between the responses of respondents who were 36 and those who were 37? What about between respondents of age 18 and those of age 65? So, in order to enhance our ability to make statistical inferences, how should ages be binned together without risking losing Predictability for the age variable? Too few bins risks losing Predictability and

[4] Assuming the respondents are taken randomly from a group of, say, age $15 - 80$.

detail, too many bins risks having statistically unreliable predictions (too few data per bin). One may imagine that with enough data one may stick with the more detailed description. This is true for a small number of variables. However, for our example of a 40 question binary survey there are over a trillion different possible completed surveys. If we have only a 1000 respondents then we still have more than a trillion possibilities about which we know nothing.

Can we *infer* anything about them? There are two basic avenues to answering this question. One is to make an informed guess for a model that one hopes will fit the data, such as a linear model in the example above. Statistical models in this sense come in many shapes and sizes with respect to the facility with which they can be used, their biases, and their degree of sophistication. They range from the simplicity of a linear regression, to the sophistication of a non-parametric estimator, such as a neural network. In this case, the effective reduction in the number of degrees of freedom of the data is associated with the reduction from the number of data points to the number of parameters in the family of models one is using. The other avenue is to let the data "speak for themselves". In this case, one does not necessarily posit a model, but rather builds up an estimation of the probabilistic relationship between the variables from the data itself, as will be illustrated in the next section. The most radical surgery in terms of coarse graining, is simply to remove some features completely. For instance, if one of the survey questions was car colour, then it would be sensible to surmise that this variable was irrelevant for predicting spending on household products say. One might then omit it from the analysis. However, it would clearly be much better to have one or more quantitative criterion by which this may be achieved.

To further consider coarse graining, we first introduce some notation that extends the possible feature values associated with a given feature. We will denote with the symbol, $*$, a sum or average over the relevant feature values. As an example, consider the conditional probability, $P(C|x_1x_2)$, to be in a class C, given specific values for two binary features x_1 and x_2. We take these values to be $1 = high$ and $0 = low$. Thus, $P(C|11) \equiv P(C|x_1 = 1, x_2 = 1)$ is the probability to be in the class C given high values for both features. Using our new notation, $x_i = *$ signifies that the variable x_i can take on *either* of the values 1 or 0. Thus, $P(C|1*) \equiv P(C|x_1 = 1, x_2 = *) = P(C|x_1 = 1, x_2 = 1) + P(C|x_1 = 1, x_2 = 0) = P(C|11) + P(C|10)$ is the probability to have a high value of feature one irrespective of the value of feature 2. This is, of course, just the *marginal* probability for C given $x_1 = 1$. An individual, modulo possible missing values, is always associated with a fully specified feature set, though more than one individual may be associated with the same feature vector. However, including in the symbol $*$ we can now consider feature vectors associated with groups of individuals. For instance, using the above two variable example, the feature vector $\mathbf{X} = \{x_1 = 1, x_2 = 1\}$ represents those individuals that have these particular feature values for both variables, whereas $\mathbf{X} = \{x_1 = 1, x_2 = *\}$ represents those individuals that have feature x_1 take value 1 and *any* value for the feature x_2. Obviously, this is a larger group than that associated with $\mathbf{X} = \{x_1 = 1, x_2 = 1\}$.

Now, whether or not a particular feature may be neglected, may depend sensitively on the object one is trying to predict or profile. For the common case of predicting or profiling class membership, two useful functions that may be used to determine how important a feature or feature value is are: ε for feature values and ε' for features. They are described by the following simple mathematical models:

$$\varepsilon = \frac{N_X(P(C|\mathbf{X}) - P(C))}{(N_X P(C)(1 - P(C)))^{1/2}} \tag{1}$$

and

$$\varepsilon' = \frac{(\langle x_i \rangle_C - \langle x_i \rangle_{\bar{C}})}{\left(\frac{\sigma^2_{iC}}{n_C} + \frac{\sigma^2_{i\bar{C}}}{n_{\bar{C}}}\right)^{1/2}} \tag{2}$$

The function ε for a given data set, $\mathcal{S}$, depends on $P(C|\mathbf{X})$, the probability that an observation associated with a given feature vector, $\mathbf{X}$, is in the class C. The data associated with those data records described by $\mathbf{X}$ are a subset, $\mathcal{S}_X$, of $\mathcal{S}$, consisting of N_X observations. It also depends on $P(C)$, which is the corresponding probability for a random sample of size N_X taken from *all* of $\mathcal{S}$. It can be used for either discrete or continuous, metric or non-metric variables. Thus, the function ε effectively measures the statistical reliability of the result of a set of observations being different to that which would be obtained from a random distribution - the null hypothesis.

As an example, for $\mathbf{X} = 1*$, ε measures the degree to which feature value $x_1 = 1$ is associated with class membership irrespective of the value of feature 2. The higher the value of ε the higher the degree of confidence that one can reject the null hypothesis that $\mathbf{X}$ is not predictive of class membership. Typically, values of $\varepsilon > 2$ indicate that the corresponding feature value, or combination of feature values, is predictive of class membership. Similarly, $\mathbf{X} = *0$ measures the degree to which feature value $x_2 = 0$ is associated with class membership irrespective of the value of feature 1. One can also consider the feature vector $\mathbf{X} = **$. In this case, $P(C|**) = P(C)$, as $**$ refers to all possible feature value combinations.

ε can be used as a useful measure of the Predictability associated with any feature value or combination thereof. When all feature values bar one, x_{ij}, are fixed at value $*$, i.e., the probabilities are summed over the different feature values for those features associated with $*$ values, then ε is a measure of the Predictability of the feature value x_{ij}. For example, for our example of a survey with 40 binary questions, then $\varepsilon(x_1 = 1, x_2 = *, ..., x_{40} = *)$ is a measure of the Predictability associated with predicting class membership for individuals who answered 1 to the first question irrespective of their answers to the other questions.

Turning now to ε', $\langle x_i \rangle_C$ is the mean of the feature x_i for class C, and $\langle x_i \rangle_{\bar{C}}$ is the mean for the complement of the class C, or some other class C'. σ^2_{iC} and $\sigma^2_{i\bar{C}}$ are the corresponding variances, and n_C and $n_{\bar{C}}$ are the number of observations in C and its complement, or other class. Naturally, in order that the difference of

the means have a significance, it is necessary that the feature x_i be metric, i.e., have an ordered notion of "big" and "small", such as age or income, but unlike race. With metric features one can think of a feature that is discriminating. For example, age as a discriminator of healthcare insurance risk - low for young people, higher for old people. However, it may be only certain feature values that are positively correlated with a certain target variable. This is even more so in the case of non-metric variables, such as race, religion, etc. The role of ε', is to determine which metric features are predictive of class membership. If $(\langle x_i \rangle_C - \langle x_i \rangle_{\bar{C}})$ is significantly different from zero and positive then high values of x_i are predictive of class membership, whereas if it is significantly negative then low values are predictive. As with ε, typically, values of $\varepsilon' > 2$ indicate that the corresponding feature is predictive of class membership.

It is important to realize that ε and ε' are neither unique, nor necessarily the best, diagnostics for determining the principle drivers. They are, however, ones that we have found to be particularly useful. ε' is clearly related to the standard Student t-test. One potential problem with this sort of diagnostic is, that if too many comparisons are made, then it is quite likely that one will find apparently statistically significant drivers just by chance, and therefore get fooled. The use of a diagnostic, such as ANOVA, that does not suffer from this potential defect is therefore potentially useful.

With ε and ε' or other similar diagnostics in hand, one can determine what are the important drivers for profiling and prediction, i.e., those features with statistically significant values of ε' and those feature values with statistically significant values of ε. Further, one can determine what are significant combinations of drivers. For instance, thinking of risk for health insurance costs, then ε for $age = high$ and $income = low$, i.e., $\mathbf{X} = age = high,\ income = low$ would be higher than ε for $age = high$ and $income = *$ or $age = *$ and $income = low$, thus showing the interaction between age and income in determining insurance risk. Finally, for those features that show very low Predictability, one may remove them from further consideration. Of course, this must be done with care, as a feature or feature value may be of low Predictability when viewed in isolation, but may not be so when viewed in tandem with another feature or feature value with which there are strong interactions.

5.3 A Simple Elementary Landscape

To illustrate concretely the concept of a Predictability Landscape we will return to the simple example already introduced in section 4.2. As an extremely common task in data mining is to establish a relationship, $P(C|\mathbf{X})$, between a class variable, C (customers most likely to be delinquent in their payments, customers who offer the most ROI, etc.), and a feature vector $\mathbf{X}$ we will consider as an example height function $P(C|\mathbf{X})$.

The determination of the posterior probabilities $P(C|\mathbf{X})$ is clearly a problem in statistical inference, hence, we require enough observations to establish the relation between C and a particular $\mathbf{X}$ with sufficient statistical reliability. $P(C|\mathbf{X})$ can be used for both profiling and prediction. In a given data set, it

can be used to calculate what features or feature values are most important for predicting class membership, i.e., to profile members of the class. In prediction, one would calculate $P(C|\mathbf{X})$ on a given set of data, where the relationship between C and $\mathbf{X}$ is known. This data set is known as a "training" set, as it is used to train the classifiers (predictors) $P(C|\mathbf{X})$ using supervised learning on known examples. Once the $P(C|\mathbf{X})$ have been determined, one can then apply it to a new out-of-sample data set, sometimes known as a "test" set, where the class membership is unknown and is inferred by assigning an individual to a class based on the values of $P(C|\mathbf{X})$ in the training sample. In other words, if an individual in the test set, whose class is unknown, is described by a feature vector $\mathbf{X}$, then they will be assigned to that class that maximises $P(C|\mathbf{X})$ in the training data (other assignment rules are possible). Obviously, as emphasised previously, a requirement for this to work is that the training set and the test set are statistically similar.

In the example discussed in section 4.2 there are two features - x_1 and x_2 - each with 6 possible values that can be arranged on a two dimensional 6×6 grid as seen in Figures 3 and 4. $P(C|\mathbf{X})$ is then a height function associated with these 36 points and we can imagine the resulting topographical landscape. Data mining is intimately associated with determining the structure of this landscape. As the height function is a proxy for the degree of Predictability there is in the corresponding variable, e.g. $P(C|\mathbf{X})$, we refer to the landscape as the "Predictability Landscape".

On a given sample, $P(C|\mathbf{X})$ will be a fixed number. However, if we pass to a different data set, due to sampling errors, $P(C|\mathbf{X})$ might be different. Testing on many different samples would lead to a distribution of points whose mean in the limit of an infinite number of samples converged to the true probability. Thus, due to the real world limitation of being restricted to a finite, often small, data set, there is some degree of uncertainty as to the height associated with a particular feature vector. This uncertainty implies that there is some non-zero probability that a point we thought to be of above average height, i.e., where feature values are positively correlated with class membership, is actually of below average height. If we have two potential classes C_1 and C_2, then we will have two associated heights. In order to find good discriminatory feature vectors, we seek those points that lead to large height differences between the two $P(C|\mathbf{X})$ for $C = C_1$ and $C = C_2$, remembering, of course, that these height differences are prone to statistical error.

So, how are characteristic features of data mining problems related to the topographical characteristics of the landscape and, further, how is this related to our choice of statistical model for data mining? We may first enquire as to what happens if there is no predictability, i.e., that it's pure random chance? In this case, the corresponding landscape is flat, i.e., a plane in the above example, and the height of this plane is $1/P(C)$, where $P(C) = \sum_{\mathbf{X}} P(C|\mathbf{X})$ is the probability to be in the class C[5]. We will refer to this as the "random chance plane". The

[5] Intuitively, the volume under the plane. In fact, this is a universal normalisation constraint when the height function is a probability.

higher a region of the landscape, the more the corresponding feature values are associated with the class C - i.e., the more positively predictive are those feature values for the class. Correspondingly, low values are associated with negatively predictive features for the class. Of course, low and high are relative terms. The natural scale at which to refer them is that of no predictability, i.e., the random distribution. So, regions are predictive if their height is significantly greater than or less than that which corresponds to the random case, which represents the null hypothesis that the feature values are not significantly correlated to class membership. The word "significantly" here is crucial because of the issue of statistical inference from a finite sample. The observed height of a particular point is subject to error, just as if we were trying to measure the latitude of the sun with a sextant from the deck of a rocking boat. The smaller our data sample the more the boat rocks! In this sense one cannot state with certainty the nature of the Predictability Landscape, but must rather make probabilistic statements about it, such as: with 95% confidence the height of the point x_{46} is higher than the height of the random chance plane hence the feature value x_{46} is predictive of class membership.

Having concluded that Predictability is associated with significant deviations in height (low or high) from a plane that represents the null hypothesis of a random distribution, we can ask what we might infer from a more detailed examination of the topography? In the above example, a predictive region could be a plateau (two-dimensional), or a ridge (one-dimensional), or an isolated peak (zero-dimensional). Also of interest, is the gradient (slope = rate of change of height) in the landscape, and the curvature (rate of change of gradient). Whether or not we will see correlations in the height between neighbouring points (i.e., smooth slopes) depends to a large extent on the nature of the features and their values. For instance, for metric variables, if they are predictive, one will typically see a gradient going from low values to high (or vice versa). For instance, for predicting those with high health insurance risk: If x_1 is age divided up into 6 categories - $0-40$, $40-50$, $50-60$, $60-70$, $70-80$, $80-90$ - then one would expect to see a gradient in $P(C|X)$ passing from lower to higher values, thus showing age to be a relevant predictive variable. Although age in this context would be a discriminatory variable, it would most probably be associated with a fairly diffuse decision boundary from a discriminant analysis point of view, say, i.e., many older people may be in relatively good health and many younger ones may have chronic illnesses. This is reflected in the fact that the *change* in gradient passing from younger to older is relatively gradual, i.e., that the curvature is small. Metric variables that are associated with sharper decision boundaries will be correspondingly associated with abrupt changes in the height of the landscape. Such changes are often naturally captured in logistic regression, where a sigmoid function is used, and where low and high values for $P(C|\mathbf{X})$ have some correspondence with low and high values for a particular feature.

Although, of course, any binary variable can trivially be regarded as metric, in that $P(C|x_1 = 1) > P(C|x_1 = 0)$, or vice versa, in general, for a non-metric variable, one would not expect to see a "trend", as there is no natural ordering

of the feature values. One could still see peaks however. If they are sharp, then that feature value has strong predictive value. Thus, if x_2 is a variable such as race, then a high value of $P(C|\mathbf{X})$ for a particular value of x_2 would indicate this value to be predictive.

The phenomena discussed above can be seen in Figures 3 and 4. First, consider the simple landscape of Figure 3. There we see that high values of feature 1 are indicative of class membership, as are high values of feature 2. Clearly, high values of features 1 and 2 simultaneously are even more indicative of class membership. Note also, that the feature values are effectively linearly independent. This Predictability Landscape would evidently be amenable to a linear analysis, such as using a linear regression.

In Figure 4 we see a more realistic representation of what can happen. Note the gradients in feature 2, indicating that high values of this feature are predictive of class membership. However, note that high values of feature 2 are not predictive for all values of feature 1, rather, only for the smallest and largest values. On the contrary, averaged over all values of feature 1, feature 2 is not particularly predictive at all. This is an important point: if we project the Predictability Landscape onto the $(P(C|\mathbf{X}), x_2)$ plane we can see the correlation between $P(C|\mathbf{X})$ and x_2 averaged over different x_1 values, thus enabling us to see whether or not there is predictability in x_2, irrespective of the values of x_1. We see that this is not the case. Similarly, if we project onto the $(P(C|\mathbf{X}), x_1)$ plane, we see that neither is x_1 particularly predictive when averaged over x_2. In fact, the only real statement that can be made, that is not contingent on the values of x_1 and x_2 simultaneously, is that there is little predictability in the entire region $1 \leq x_1 \leq 4$, $0 \leq x_2 \leq 5$, apart from an isolated peak at $x_1 = x_2 = 3$, the latter being like a "needle-in-a-haystack". Also, note that, even in the cases where there is a Predictability gradient, as in high values of feature 2 when feature 1 is either high or low, the form of this gradient can be quite distinct. For $x_1 = 0$ the gradient is smooth while for $x_1 = 5$ there is an abrupt change at $x_2 = 3$. This has implications for how discriminating these variables are.

Qualitatively, the chief difference between the Predictability Landscapes in Figures 3 and 4, is that the former is uni-modal (only one local maximum or minimum of the function), whereas the latter is multi-modal (more than one local optimum). This fact has important consequences for how easy it is to search for Predictability on the landscape, and what are the optimal models to use to detect it. Note that, although we are imagining the Predictability Landscape in the context of a particular target variable - $P(C|\mathbf{X})$ - similar considerations will hold for any target function $F(\mathbf{X})$ that depends on the feature vectors. For instance, $F(\mathbf{X})$ might be the expected lifetime ROI of customers as a function of $\mathbf{X}$.

It is crucial to realize that the topography depends sensitively on the particular target function. As an example, consider once again the single feature - age. For a health insurance company predicting healthcare risks one would expect that after the infant years there is an increase as a function of age with a slope that steadily increases as well. However, now consider a beer manufacturer in-

terested in gauging alcohol consumption as a function of age. In this case, there would be a pronounced peak around college age! Even more complex, consider the probability that a person will shop for toys as a function of age. On average, this will show a tendency towards a double peaked structure with one peak - the "parent's peak" - being associated with the most probable ages for being a parent, and another - the "grandparents' peak" - associated with the most probable ages for being a grandparent. Thus, we see how the topography can show quite different features for different target functions as a function of the same feature variable. This latter double peaked structure is an example of multi-modality and is not a pathology but something present in almost all non-trivial data mining problems.

So, we see that successful data mining is really associated with determining the topographical features of a landscape. This is quite intuitive in the case of a low dimensional problem. However, imagine now the situation where instead of two dimensions we have 200 (difficult to visualise!). What we are trying to do is infer the topography from a set of values of the height function. If we have many values for each point in the space then we can proceed relatively straightforwardly. However, if this is not the case, then it is not possible to determine the height with precision. In fact this is almost never the case. As inevitably we are modelling probabilistic processes, the height function associated with the landscape is statistically inferred with some degree of uncertainty. The more data we have available for a particular feature vector, the more precisely we can infer the corresponding height. However, mistakes can be made. We may think that a point on the landscape is high. It may be that this is a pure sampling effect though. If we passed to another data set the corresponding point may be relatively low. In general, we must address the question of: How does the topography change when we go to a different data set? Did a valley or a hill disappear? Did an uphill gradient turn into a flat plane? Also, how does the topography change as a function of time? This susceptibility to error happens when the estimated Predictability Landscape, as determined from one set of observations, does not generalise to another data set. This is, of course, the potential pitfall of all statistical estimators - "over-fitting". Here we see it posed in the geometrical language of the Predictability Landscape.

As mentioned, a crucial aspect of this topographical point of view is that certain data mining models are more suited to a particular kind of topography than others. The first thing to be interested in is the gross topographical properties of the landscape: Is a particular region high or low? Are there many peaks and valleys, or is the landscape relatively smooth? Are there trends - does the ground generally rise in a particular direction? The important thing about determining the gross topographical features is that it is statistically more likely that we will be able to infer them with a higher degree of confidence. For instance, in a linear regression, we are required to fix only N gradient parameters. If the total number of data points is much greater than N, then they can be determined with precision. However, the bias associated with the model, linearity in this case, might be quite inappropriate, especially if there are non-linear interactions

between features. On the other hand, when we go to consider a more detailed picture of the Predictability Landscape, although the model bias may be reduced due to the introduction of more parameters, it is easier to make an error due to over-fitting the data.

The grosser structure of the Predictability Landscape can be teased out using coarse grained models. These can be models that have a strong bias - such as a linear regression - or models that have little or no bias. In the present case, with $P(C|\mathbf{X})$ as predictability measure, a fully specified feature vector with no $*$ represents a point in the space, e.g. $P(C|x_1 = 3, x_2 = 5)$. If we include an $*$ in $\mathbf{X}$ however, then the resultant $\mathbf{X}$ represents a line rather than a point. The associated $P(C|\mathbf{X})$ represents a sum along a line rather than a point on the Predictability Landscape. For example, $P(C|x_1 = 3, x_2 = *)$ represents the sum of the heights along the line $x_1 = 3$. In the case of more features, these notions generalise - the more $*$, the higher the dimension of the hyperplane represented by $\mathbf{X}$. For instance, for three features $\{x_1, x_2, x_3\}$ taking integer values between 0 and 9, then $\{x_1 = 3, x_2 = 7, x_3 = *\}$ is a line parallel to the x_3 axis that passes through the point $x_1 = 3$, $x_2 = 7$, while $\{x_1 = 3, x_2 = *, x_3 = *\}$ is a plane perpendicular to the x_1 axis and passing through the point $x_1 = 3$.

Before leaving this section, we wish to point out two more important characteristics related to earlier discussions: Firstly, that the Predictability Landscape can be time dependent. In 3.2 we pointed out that the statistical distribution that underlies the data may be time dependent. This implies that, under such circumstances, the Predictability Landscape itself will change. Thus, a point that was of above average Predictability over a certain time period, i.e., was above average height, may later decrease until it is of only average height, revealing that the associated feature values are no longer predictive. The second point to emphasise, is that there exist barriers to how much Predictability can be gleaned from a problem, as discussed in section 3.4. This means that there is a limit to the height of the points in the Predictability Landscape that cannot be overcome.

6 Data Mining Models and the Predictability Landscape

We have emphasised that data mining can be viewed as the search for Predictability. We now wish to focus on the interplay between different data mining models and the topographical features of the Predictability Landscape. As making an exhaustive analysis of the different model classes is beyond the scope of this article, we will restrict our attention to some of the model classes that we have found to be most useful.

6.1 Data Mining Models as Topographic Templates on the Predictability Landscape

So, data mining is the discovery and analysis of patterns in data. These patterns are relations, $F(\mathbf{X})$, between input, $\mathbf{X}$, and output, F, variables. Any of

these relations can be visualised in terms of a topographical landscape - the Predictability Landscape - with the output variable viewed as a height function. What the most interesting features of this landscape are depends on the particular problem being studied. Often, what is of interest is to find the high points of the landscape, i.e., the mountain peaks. This is particularly true when the output variable is a probability for instance. It would also be true if the output variable were a cost or ROI-type function, where we would be looking for those variable combinations that lead to the most ROI. On the other hand, if one is looking for discriminatory variables, one might be interested in regions where there was a sharp increase in height passing from one region to another. Further subtleties are engendered by thinking about the size of the regions that are of a given type. For example, if a mountain peak is isolated, as in Figure 3, then it might well be that few individuals are associated with the feature values of that peak. Once again, with Figure 3, only those individuals with feature $1 = 3$ and feature $2 = 3$ are associated with the peak there. Imagine now, however, that the peak had been a broad plateau, encompassing the area $1 \leq$ feature $1 \leq 4$ and $1 \leq$ feature $2 \leq 4$. In this case, the number of individuals associated with this high region is likely to be greater. So, not only the height of the region is of interest, but also its size and shape.

Remember, in real-world data mining, we are usually only given a small number of data points, relative to the total number of possible ones, for which we know the relation $F(\mathbf{X})$ between a set of known values of F and $\mathbf{X}$. This is often known as the training set. From this set of observations, the job then is to *infer* the value of F for values of $\mathbf{X}$ that we have not previously seen. This obviously requires an *interpolation* which,preferably, should be most faithful for those regions of the Predictability Landscape of most importance, e.g., the high regions, etc. A data mining model is a template that we "fit" to the observed landscape. We judge it by how well it describes the landscape. The template then offers a way to interpolate from the known height at known values of $\mathbf{X}$, to infer the height at other values of $\mathbf{X}$. This is true even in the most elementary cases, as in Figure 1, where to infer the ROI for a spending of \$2.5 million in advertising one must interpolate the relation between spending and ROI to other values than those for which there exist observations. There are, of course, many choices of template. Our two VPs chose different templates - one a linear function and the other a sixth-order polynomial. In Figure 3 the form of the landscape argues for a linear function of the two features as being suitable.

We have alluded to the fact that there are simple, parametric models, such as linear or logistic regression, and that these models have a strong bias. Such models, unless the bias captures something inherent in the data, can yield quite poor results. However, due to their simplicity there is little chance of over-fitting the data. More complex parametric models, such as a non-linear regression, or non-parametric models, such as neural networks, although offering greater flexibility, risk over-fitting. The noisier the Predictability, i.e., the more dispersion in its values for a given feature vector, the more risk that the model may over-fit the data. In general, it is better to avoid overly complex or sophisticated models

unless one has ample experience at recognising and avoiding over-fitting, as the latter depends subtly on various parameters, such as the fundamental dispersion in the Predictability measure and the sample size. The latter, which gives rise to sampling error, can be ameliorated by taking more observations, the former cannot. In this general sense, the more data you have to work with the more confidence one can have in using a model with more parameters, or degrees of freedom. Speaking loosely - for a given sample size, the more parameters one has in a model, the less "data per parameter" one has with which to statistically determine the parameters.

Similar considerations also apply in models where one estimates the probability distribution of the Predictability measure directly from the data itself, e.g., where one constructs $P(C|\mathbf{X})$ directly from the data. As we have emphasised, the number of possible feature vectors, $\mathbf{X}$, is inevitably much higher than the number of data records. Consequently, typically, for the vast majority of feature vectors we have no observations from which to make statistical inferences, while for those observations we do have, we are almost inevitably making inferences from a sample of one! We have advocated the use of coarse graining in order to ameliorate this conundrum, as higher dimensional objects in the Predictability Landscape, such as lines, planes, etc. will be associated with more observations. For example, with our 6×6 grid, a line contains 6 points hence, if on average there are 10 observations per grid point, then the number of observations that can be used to estimate $P(C|x_1*)$, for example, will typically be 60. Obviously, this gain in statistical reliability increases as more and more features are coarse grained, i.e., we consider higher and higher dimensionality hyperplanes. However, after determining with better statistical confidence these coarse grained measures, one has to face up to the problem of how to construct fine grained predictions. For example, given $P(C|x_1*)$ and $P(C|*x_2)$, how does one determine $P(C|x_1x_2)$? Obviously, there is no answer to this question without further assumptions or approximations. However, one very robust and powerful methodology that we have found to be useful in that regard, in both academic and real-world problems, is that of Naive Bayesian classifiers

6.2 Naive Bayesian analysis

This is a computationally cheap, easy to implement classifier that can be shown to be the optimal method of supervised learning if the feature values are independent given the class of the example. Of course, the opposite is also true: when there are correlations, then it cannot be the optimal method. However, there is ample evidence to show that naive Bayesian learning is remarkably effective in practice, and difficult to improve upon systematically [33].

The object is to construct the posterior probabilities $P(C|\mathbf{X})$ from more statistically reliable coarse grained entities. To achieve this using "naive" Bayesian analysis, we start with Bayes rule

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})} \tag{3}$$

which relates the posterior probability to the prior probability, $P(C)$, for the class C, and the likelihood function, $P(\mathbf{X}|C)$. For this to be useful one must still compute the right hand side. In the case where the features or feature values are independent however, we may write

$$P(\mathbf{X}|C) = \prod_{i=1}^{N} P(x_{ij}|C) \tag{4}$$

where $P(x_{ij}|C)$ is the marginal probability of finding the value x_{ij} for the jth feature value of feature i given class C, and the product is taken over the N features.

If we imagine a fixed number of feature values per feature, 10 say, then, for a given non-coarse grained feature vector, the probability of an individual in the population being described by this feature vector is $1/10^N$. So, generally, as emphasised, we have no statistics at all on most of the possibilities. If however, we consider the marginal probability $P(x_{ij}|C)$, where only one feature and its corresponding feature value are specified, then the probability of an individual being described by this feature value is only 1/10. Under these circumstances, one expects to find a much more statistically reliable sample to test what are the most important single variable drivers. As one would typically expect sampling errors to scale like $1/N^{1/2}$, one can see that the typical sampling error associated with the marginal probabilities, i.e., the coarse grained objects, is much lower than that associated with a fully specified feature vector. This is probably the principal reason why naive Bayesian classifiers work so unexpectedly well, especially on small to mid-size data sets, even when it is known that the independence assumption for different features is incorrect. This is very much a "swings and roundabouts" argument: what we gain on the interacting "swings", i.e., taking into account statistical dependencies between features, we lose on the "roundabouts" of statistical uncertainty, i.e., the number of data is less. Thus, for example, $P(x_1x_2|C)$ may be a more faithful representation than $P(x_1*|C)P(*x_2|C)$ when x_1 and x_2 are dependent. However, the statistical error due to sampling errors associated with $P(x_1x_2|C)$ is higher than that associated with $P(x_1*|C)P(*x_2|C)$.

The effective variance reduction implicit in the use of Naive Bayesian classifiers means that they are quite robust for supervised learning when passing from a training set to out-of-sample data, i.e., they generalise well and are not that prone to over-fitting. Thus, if our goal is to predict class membership on a given small or medium sized data set then Naive Bayesian classifiers are a very good choice of model. However, if our task is to profile and identify those feature values which are particularly correlated with respect to class membership then Naive Bayesian classifiers are useless due to their neglect of feature-feature interactions.

7 A Multi-perspective Agent-based Approach

Much that we have discussed up to now, has been to show that there is no "magic bullet" data mining model. The multi-objective requirements of real-world business problems, combined with the realization that any data mining model has its own particular imperfections and fallibilities, leads one to conclude that a single model associated with a single Predictability measure is quite inadequate. Different models are better for different tasks. For example, considering profiling and predicting - Naive Bayesian classifiers might be very good for predicting class membership in sparse data sets, but are incapable of discovering potentially important non-linear dependencies between different features or feature values. We can summarise these thoughts by stating that data mining really requires a "multi-perspective" approach. Perspective, here, could refer to a task class, such as profiling or predicting; to a particular Predictability measure, such as $P(C|\mathbf{X})$ or some other function $F(\mathbf{X})$; or to a particular statistical model. Some important types of perspective are:

- Variable Classes - for example, socio-demographic variables versus product purchasing history; or medical condition history versus medical cost history
- Time scales - for example, recent product purchases versus long term purchasing behaviour; or recent sickness burden versus long term sickness burden
- Spatial scales - for example, product purchasing behaviour in a particular zip+4 versus nationally; or average sickness burden in a particular workplace versus in an entire industry
- Statistical models - for example, naive Bayesian classifiers are robust, but cannot describe non-linear dependencies between variables, whereas neural networks are useful non-parametric estimators but highly non-transparent

Such a multi-perspective approach is very much the methodology adopted by humans when faced with a complex task. Different people are more skilled in one task than another. Also, different people bring different insights, experiences, etc. Consensus decisions reached among a group often have more validity than those reached by an individual, in terms of both reducing error as well as individual bias. In order to capture the flavour of this approach, we will work within an "agent"-based paradigm. Agent-based systems have become more widespread in recent years [34] in many different areas. There are several concrete technical definitions of what constitutes an agent. However, here we will adopt a much looser definition, so that a multi-perspective approach to data mining can be naturally discussed within an agent-based framework. We will take the point of view that:

- *An "agent" is a model that provides an opinion, perspective, profile or prediction based on input data.*

At the heart of this paradigm is the anthropomorphic inspiration, alluded to above, of a "team of experts", an agent being an "expert". Human experts may

specialise in analysis of certain geographical regions, or certain time periods, or certain subfields of knowledge. They learn and adapt in the light of new information. Training an expert may take a long time, and require the examination of many candidate experts. Decisions will often be taken at a corporate level - pooling knowledge and experience. In terms of agents:

1. One can understand "expert" here to be an agent that represents a statistical model, artificial intelligence technique or other functionality.
2. Agents may be adapted to particular temporal and/or spatial scales as appropriate, or be trained using a particular subset of data.
3. In dynamic problems, agents adapt in the presence of new data.
4. Discovery of useful agents generically requires an intelligent search through the space of agents.
5. Agent profiles and predictions are combined without consensus to give a multi-perspective point of view, or with consensus to provide an improved uni-perspective viewpoint.

Agents can be of different functional classes, tasked to help on different aspects of the data mining problem. For instance, in section 5.2 we introduced the quantities ε and ε' that measure the relevance of particular drivers, or combinations of drivers, with respect to some target function. Thought of as agents, these two quantities represent the agents' (quantitative) "opinions" about what are the important drivers. They are thus important in the aspect of coarse graining associated with dimensional reduction, i.e., identifying the subset of predictive variables from which predictions and profiles will be formed. One may also subdivide this feature detection class of agent into subclasses associated with different classes of data. For instance, one may think of an agent of this type specialised to use only socio-demographic data. This agent would then be appropriate for forming opinions, for example, about potential new customers, where one has access only to socio-demographic census-type data. Another subclass may be associated with only medical data if one is interested in giving a medical rational for why someone predicted to be a high cost individual in terms of healthcare costs should be put into care management.

The pattern should be clear - for the many different potentially orthogonal, potentially complementary, potentially parallel, simultaneous objectives in a data mining project, one or more classes of agents may be assigned to a particular objective. Additionally, one may include as agents one's own favourite statistical models, or some other "benchmark" model. For instance, maybe a linear regression has been found to give good results for a particular aspect of a data mining problem. One would then naturally include in this model as an agent, thus leveraging previous analysis, while at the same time generating and examining other agents, comparing them to the linear regression and then integrating the results from the different models together.

Passing from feature detection to prediction, one can now think of agents that represent different prediction tools. To give some examples: one agent may represent a linear regression with associated slope and intercept parameters, optimised by fitting to the data set under discussion, but only on those variables

pre-selected by the feature detection agents. Another may represent a naive Bayesian classifier, while yet another may represent a neural network, once again, taking as inputs only those variables pre-selected by the feature detection agents.

Further on in the process, one may then think of agents that represent certain actions or strategies that are suggested in the light of the knowledge gleaned from the feature detection and prediction agents. For instance, in the case of Customer Resource Management such an "action agent" might represent the passing of a list to a human agent of those customers most likely to defect from the company in order that efforts may be made to guard against such churn.

Clearly, there is a hierarchy and taxonomy of agents, just as there is in any human organisation. The degree to which this hierarchy may be automated depends on the problem at hand and the type of agent. Human participation can occur, or may be necessary, at each step of the process. In this sense the autonomous agent/human interface can range from being as simple as a human agent utilising a linear regression (prediction "agent") and interpreting the results, to that of an autonomous system such as in ATi's portfolio management system Adaptrader where, in principle, no human intervention is needed whatsoever, and teams of artificial agent "traders" do all the trading.

To get a more concrete feeling for how an agent is represented mathematically, consider an agent that represents a naive Bayesian classifier. Mathematically, the agent is represented by

$$P(class = C|\mathbf{X} = \mathbf{x}) = \prod_i P(X_i = x_i|class = C)P(class = C)/P(\mathbf{X}) \qquad (5)$$

which is a classifier that links a set of feature values, $\mathbf{X}$, associated with a potential customer, to a target class C. For example, C could represent potential purchasers of a car insurance policy and $\mathbf{X}$ a set of socio-demographic features with individual component values X_i (e.g. x_1 = purchasing power class high, x_2 = home owner). $P(X_1 = x_1|class = C)$ is then the conditional probability that, a purchaser of car insurance was in the high purchasing power class, while $P(X_2 = x_2|class = C)$ is the probability that a purchaser was a home owner. Intuitively, they represent agents' "opinions" about the relationship between buying a car insurance policy and purchasing power class and being a home owner respectively. Then, $P(class = C|X_1 = x_i, X_2 = x_2)$ is the conditional probability that a high earner and home owner will purchase a car insurance policy. This multi-perspective classifier now represents the "net/consensus" agent opinion about how likely it is that a potential customer with high purchasing power class who is a home owner will buy a car insurance policy.

Several questions are naturally raised by this multi-perspective point of view: How does one measure the performance of an agent? How does one generate agents? and, finally, how does one integrate together the different opinions, predictions and actions suggested by the agents? On the question of agent performance, from a Darwinian perspective, one may fruitfully think of an agent's performance as being a measure of the agent's "fitness". This is a measure of the utility of the agent, e.g. sensitivity, specificity, accuracy, etc. Once assigned

a fitness, one may think of an environment within which agents compete against one another. Of course, this competition is hierarchical and modular. It does not make any sense to compare the performance of agents that are predicting who to target for a new marketing campaign with those chosen to determine the most likely targets for a cross-sell among the company's current customer base.

In terms of generation of agents, one may do this selectively "by hand", which to a large degree would be the traditional data mining approach, or one may use some automated technique. Of course, the space of potential agents is very large, and hence one must determine what part of the space one is going to search. A human designer will utilise experience and intuition to limit the search to a very small domain, whereas an intelligent automated search technique may be capable of searching a much larger number of possibilities. In this vein, Evolutionary Computation techniques, such as Genetic Algorithms [27, 28], Genetic Programming [31, 32] and Evolution Strategies [28] can be of great use. There are a large number of search heuristics that can be useful, ranging from the simplicity of local gradient descent type methods to much more complicated techniques, where the risk of failure is much higher when trying to implement them without adequate knowledge and experience.

Finally, how does one integrate information from among different agents? There are many different ways of achieving this, and final results may be quite sensitive to how it is done. A common integration task occurs when more than one agent opinion/perspective is included for a particular task. Hence, a means of combining such opinions/perspectives must be sought. For instance, in the context of classification, an individual may be classified by specifying a "consensus" rule. Such a rule assigns a score to an individual which measures the probability that this individual is in the relevant class and is derived from the agents associated with, or activated by, this individual. For example: If the list of classifier agents is $P(class|X = 11)$, $P(class|X = 1*)$ and $P(class|X = *0)$; an individual described by the feature vector $X = 10$ is associated with, or activated by, both the second and third classifiers, as the feature vector $X = 10$ is contained in both $1*$ and $0*$. On the other hand, an individual, represented by the feature vector $X = 01$, activates none of them. Some useful consensus rules are:

- Winner-takes-all - here, an individual's score is the fitness of the highest ranked agent activated by that individual. This consensus rule takes only the opinion of the highest ranked agent
- Average - in this case, an individual's score is the average of the fitnesses of the agents activated by that individual. Here, one takes a weighted consensus among the activated agents so that the opinions of all activated agents are taken into account
- Match - where, an individual's score is determined by the number of agents activated by that individual. In this case one takes an unweighted consensus among the activated agents

8 The Multi-perspective Approach in Action

We may take as a model problem to illustrate some of the points we have made an "academic" (though taken from a real-world) problem (used for the COIL Challenge 2000 competition [35, 36]). The original competition goal was to predict who would be most likely to buy a caravan (mobile home) insurance policy from a group of customers of the company. In particular, to select a subset of 800 from the test set that are considered most likely to purchase a policy. The potential applications are clear: cross-selling to actual customers of the company, identifying potential new customers, etc. The information associated with each of these potential target classes would, of course, be different, information on potential new customers most probably being restricted to census-type socio-demographic data.

The data consisted of 9322 individual customer records consisting of:

- 86 features
- 43 socio-demographic features (at the "zipcode" level, e.g. percentage of customers of high purchasing power class in that zipcode)
- 43 product purchasing attributes (e.g. number of car insurance policies bought).

The data from 5322 individuals were used as in-sample training data, while that of the remaining 4000 was reserved for out-of-sample testing. The number of mobile home insurance policy purchasers in both the training and test sets was about 6%. We take as a concrete example - $P(C|\mathbf{X})$ - the probability to purchase a mobile home insurance policy (to be in the class of purchasers) given a feature set $\mathbf{X}$. The number of possible feature vectors, n, in this case, is $> 10^{80}$ (the approximate number of atoms in the universe!). The number of training data points is only of the order of 10^4, which means that we have no knowledge at all about the vast majority ($10^{80} - 10^4 \sim 10^{80}$) of possible feature vectors! In other words we cannot with any confidence predict whether a particular individual will or will not purchase a policy as every individual is almost unique. In other words, there is no way to assign an individual to a class when an individual of that type has never been seen before. One can't do supervised learning without supervision!

To ameliorate this grave problem we can try to implement the philosophy espoused so far - identify what are the most relevant features, or feature values, and then try to classify based on this reduced subset of features. We would like to choose those feature vectors that are most predictive of class membership, but without being able to calculate $P(C|\mathbf{X})$ we do not know a priori which ones are predictive, so we have a Catch 22 situation. A common way to circumvent this problem, is to forget about trying to identify predictive feature vectors and to go to the simpler task of trying to identify predictive features or feature values individually. This is a very significant coarse graining.

8.1 Uni-perspective Profiling

To proceed, we can first look for predictive feature values or predictive features using, for example, ε and ε' as discussed in section 5.2. As we are only exam-

ining individual drivers we term this a uni-perspective approach as we are only incorporating information from one particular perspective. Below, in Table 6 we see the most significant single-variable drivers of mobile home insurance policy purchasing based on ε. "Number of customers with driver" is that number in the training (in-sample) data set that correspond to the associated feature value, while "Number of purchasers with driver" refers to that subset of customers with that feature value that also bought a policy. % refers to that percentage of customers with a given feature value that bought a policy. The higher this number the more likely the customer is to have a mobile home insurance policy.

Profile Driver	ε	No. Customers with driver	No. Purchasers with driver	%
Auto insurance contribution $500-2500	10.81	2319	262	11.3
Fire insurance contribution $100-250	9.36	1226	151	12.3
Boat insurance policy	7.69	31	12	38.7
High Purchasing power	7.49	474	67	14.1
Middle class families	7.04	339	51	15.0
Driven Growers	6.78	502	66	13.1
2 Auto Insurance Policies	6.27	246	38	15.4
1 Auto Insurance Policy	6.07	2712	237	8.7
Third Party Insurance Contribution $25-50	5.83	2128	191	9.0
1 Third Party Insurance Policy	5.37	2334	201	8.6
Social Security Insurance Contribution	5.23	81	16	19.8

Table 1: Most important uni-perspective drivers for mobile home insurance purchasers

What can we glean from these drivers? First, notice the very high percentage of boat insurance policy owners who have also purchased mobile home insurance. This is not the top ranked driver in terms of ε however, due to the relatively small number of boat policy owners - only 31 in the whole sample. Notice that there are only three socio-demographic factors involved - "High Purchasing Power", "Middle class families" and "Driven Growers" - all of which are somewhat related - and that indicate that the more economically advantaged are likely to purchase a policy. This is confirmed by noticing that two car families are nearly 80% more likely to purchase a policy than those with only one car. Note that, generically, socio-demographic data tends to be less predictive than other data types, such as product purchasing behaviour. From this we can determine a preliminary profile for mobile home insurance purchasers:

- Middle class
- High income
- "Risk Averse" (have multiple types of insurance policies)
- Auto owners
- Boat owners

Thus, as one might expect, mobile home policy purchasers tend to be wealthier and are somewhat more "risk averse", i.e., they tend to have multiple types of insurance policy.

In terms of targeting prospective purchasers of a particular product - i.e., those who are *more* likely to buy the product - it is also useful to be able to identify those *less* likely to buy the product. In the present case, we examine in Table 6 those drivers which are most associated with those least likely to purchase a mobile home policy.

Profile Driver	ε	No. Customers with driver	No. Purchasers with driver	%
50% of people in area have no auto	-3.85	587	13	2.2
Auto insurance contribution $250-500	-3.86	613	14	2.3
No fire insurance policy	-4.11	2666	109	4.1
No third party insurance policy	-4.37	3482	147	4.2
Fire insurance contribution $25-50	-4.74	535	6	1.1
No auto insurance policy	-7.75	2845	72	2.5

Table 2: Most important uni-perspective drivers for non-purchasers of mobile home insurance

Here we see that those having no auto are much less likely to purchase a mobile home policy. Also, we see a correlation with purchasing other insurance products. There are at least two potential explanations for this - that the associated customers cannot afford insurance products, though interestingly, there are no particularly predictive socio-demographic features to confirm this - or that these customers are less risk averse.

8.2 Multi-perspective Profiling

Above, we viewed the problem from a range of single perspectives, i.e., single variables in this context. If there were no correlations or non-linearities between variables, then this might be adequate. However, it is also important to see, for instance, whether there are combinations of drivers that are more predictive together than separately. As mentioned, the problem is that, generically, there are a vast number of potential combinations of drivers. In Table 6 we see the most important combinations, as found by an evolutionary search using a genetic algorithm.

Profile Driver	ε	Customers with driver	Purchasers with driver	%
Auto insurance contribution \$500-2500 AND Fire Insurance policy	12.61	1315	187	14.2
Auto insurance contribution \$500-2500 AND Third party insurance contribution \$25-50	12.26	1070	159	14.9
Auto insurance contribution \$500-2500 AND No Farmers in area	11.37	1714	214	12.5
Auto insurance contribution \$500-2500 AND Driven Grower	11.26	215	52	24.2
Auto insurance contribution \$500-2500 AND No contribution to agricultural insurance	11.15	2240	259	11.6
Auto insurance contribution \$500-2500 AND No rented housing in area	11.01	407	77	18.9

Table 3: Most important multi-perspective drivers for purchasers of mobile home insurance

Interestingly, all combinations have as one component a high contribution to Auto Insurance, which was the principle uni-perspective driver. Notice though, that in combination with other less predictive drivers the result is much more predictive than either of the component perspectives. Thus, being a Driven Grower *and* having a high Auto Insurance contribution makes one twice as likely to be a purchaser of a mobile home policy as having either of these characteristics separately. This significant non-linearity is not something that could have been determined using a linear analysis. Nor is it something that could have been detected using a naive Bayesian analysis. This inherent non-linearity also implies that if linear discriminant analysis had been used to predict those who are most likely to purchase the results would have been quite suboptimal. We also see that predictability is somewhat enhanced by considering a more urban environment - i.e., where there are no farmers, or no contribution to agricultural insurance. These combinations are consistent with the uni-perspective profiles found in section 8.1, i.e., that mobile home insurance policy purchasers are risk averse - owning various other forms of insurance policy - and being higher income individuals, as evidenced by the fact that they live in areas where housing is by 100% owner-occupiers.

Both uni-perspective and multi-perspective pictures can be combined together to generate richer, more predictive profiles. Uni-perspective profiles will tend to be less predictive than their multi-perspective counterparts, but, being less restrictive, are usually applicable to a larger sample. For instance, the class of Driven Growers is larger than the class of Driven Growers with a high Auto Insurance contribution. Additionally, which perspectives are applicable depends very much on the group one is trying to predict. For instance, in the present case, if one wishes to predict prospective purchasers of mobile home insurance from outside the company's present customer base, then one is unlikely to have

access to other than socio-demographic data. Hence, a suitable profile must be gleaned from only the socio-demographic variables. On the other hand, if one is trying to up-sell across the current customer database, then it is appropriate to use all variables. This is another area in which the multi-perspective nature of the problem may come into play - identifying potential up-sells within the company's database versus identifying prospective new customers from outside the database.

8.3 Multi-perspective Predictions

As an illustration, we will consider two related classes of prediction agent: Naive Bayesian classifiers and Posterior Probability Classifiers, where, in the latter, the posterior probabilities are calculated directly from the data. Thus, the agent ranks customers on their probability to purchase a mobile home insurance policy based on classifiers of the form $P(\text{buy policy}|\mathbf{X})$, where $\mathbf{X}$ is a set of feature values associated with zip-code based socio-demographic factors and previous products purchased. In the case of Posterior Probability Classifiers, as the dimensionality of the feature space is high, a genetic algorithm was used to intelligently search for fit classifiers.

Different prediction agents were also produced by concentrating on different variable classes as potential feature sets. For example, agents that considered only socio-demographic variables; agents that considered only variable class information associated with previous products purchased, i.e., ignoring socio-demographic data; or agents that used previous purchasing behaviour combined with information gleaned from only the most predictive socio-demographic variables, as determined by the profiling agents.

It is also instructive to show how Predictability and performance are related to different types of consensus rule. In Table 4 below, we show the results for some different variable classes and consensus rules. All %s here refer to the percentage of correctly identified purchasers in the (out-of-sample) test set. Remember that about 6% is what would be found if potential purchasers were picked at random.

Consensus type	Average	Match	Winner
All variables	7.0%	13.2%	11.6%
Demographic variables omitted	7.9%	13.4%	11.7%
Most predictive demographic variables used	7.3%	14.1%	12.5%

Table 4: Relative Performance (Sensitivity) of various consensus functions and different variable-class perspectives for identifying the 20% of customers most likely to purchase mobile home insurance

Note that performance tends to improve when socio-demographic variables are omitted relative to keeping all of them. In other words, better to have no socio-demographic information at all rather than include all of it. This shows the importance of the feature selection step before developing a prediction model.

However, by determining the most predictive socio-demographic variables using feature selection agents (e.g. using ε and ε' from section 5.2) Predictability can be enhanced. The implication of this for model building using only socio-demographic data is that a model built only on the four most predictive socio-demographic variables is superior to one built on all 43. So, variable class not only enters due to constraints, such as whether one has access to the data or not, but also due to questions of Predictability. Turning now to the different consensus rules: we see that taking an average of the agents opinions leads to poor results - only moderately better than random choice. Just using the opinion of the fittest agent led to much better results. However, the best results were associated with just counting the number of agents that shared the same opinion about whether or not a customer was likely to purchase a policy.

8.4 Making Predictions and Profiles Actionable

We have emphasised that good data mining needs to keep in mind the business goals that spawned it and, in particular, to have in mind as high a degree of actionability as possible. Too often, formulaic segmentation analysis in terms of exotic sounding socio- or psycho-demographic segments results in little or no actionable analysis. Here, we will illustrate with a simple scenario how the analysis above could be made actionable in terms of increasing a company's ROI.

- Company goals: Company wants to increase the ROI associated with selling to its present customer base. It will test the ability of multi-perspective prediction and profiling technology to improve the accuracy of targeting those customers most likely to purchase a mobile home policy. The test sample is a set of 8,000 of the 40,000 registered customers of the company in a large Midwestern city. A comparison between the results from direct sales calls to a random non-targeted sample will be compared with the equivalent results from a targeted sample found from a multi-perspective approach

- Assumptions:
  1. Number of actual customers contacted = 8000 = 20% of total actual customer data base for city of interest
  2. The probability that an untargeted direct sales call to present customers will result in a sale = 6% (number taken from section 8)
  3. The probability that a targeted direct sales call to present customers using multi-perspective approach = 12.5% (number taken from results of section 8.3, Table 8.3)
  4. Average cost of an unsuccessful call = \$5.00
  5. Average cost of a successful call = \$10.00
  6. Average income from sale of mobile home policy = \$300.00

- ROI predictions
  1. Average ROI per customer with untargeted calls = $0.06 \times (\$300 - \$10) - 0.96 \times \$5 = \$12.60$

   2. Average ROI per customer with targeted calls = $0.125 \times (\$300 - \$10) - 0.90 \times \$5 = \$31.75$
   3. Total expected ROI with untargeted calls = $8000 \times \$12.60 = \$100{,}800$
   4. Total expected ROI with targeted calls = $8000 \times \$31.75 = \$254,000$
- ROI results
   1. For the untargeted sample the 8000 random sales calls resulted in 457 sales leading to a total ROI of $457 \times (\$300 - \$10) - 7543 \times \$5 = \$94,815$
   2. For the targeted sample the 8000 targeted sales calls resulted in 811 sales leading to a total ROI of $811 \times (\$300 - \$10) - 7194 \times \$5 = \$199,220$

Thus, we see through the above analysis that being better able to target prospective purchasers of mobile home insurance policies leads to an increase in ROI of over 100%. This, of course, just represents one simple example of how profiles and predictions may be turned into actionable business intelligence.

9 The Take Home Message

In this section, we will try to distill some of the key points that emerge from the discussions in this paper into a more succinct format, remembering that much of the detail is found in the previous sections. The key strategic steps in a data mining project are:

1. Specification of business goals and priorities
2. Data collection
3. Data analysis
4. Implementation of actionable business intelligence

Of course, the organisational elements and personnel involved in each step will most probably be different. However, to maximise any added value from the project, one has to have an adequate effective integration and communication of understanding and expectations between these different organisational elements. Too often we have seen a scenario of the following type played out:

1. Management set goals without due regard as to the technical feasibility of collecting the necessary data, or whether such data can be successfully mined
2. Data collection is carried out using "out-of-the-box" instruments and metrics without understanding the company's business goals, the potential degree of Predictability in the data or the actionability of anything that may be gleaned from its subsequent analysis
3. Analysis is done using a single, presumed "magic bullet" ("best-of-breed" is a common terminology) model
4. Groups/clusters are identified that have low predictability and cannot be readily targeted

The net result is disappointed expectations on all sides. The data collectors, usually an outsourced company, are disappointed that the analysts cannot find any Predictability in their data. The analysts, more often than not a different

outsourced company, are disappointed that they can't intuitively understand the data, as they played no role in choosing what type of data to collect. Finally, management is disappointed that after a substantial investment very little of any substance emerges from the entire project, and what does emerge is not readily actionable. A key problem is that the success of each stage intimately depends on the others. The reality however, is that they are often treated in isolation by completely separate groups.

A successful data mining project will involve the following steps, there being two principle elements - Plan to Predict and constructing and understanding the Predictability Landscape using a Multi-perspective approach.

1. Plan to Predict
   (a) Establish business objectives and priorities
   (b) Communicate them to the data collectors and the data modellers. These goals need to be translated into requirements on which Predictability measures to use and modelling priorities, such as precision, interpretability, transparency, etc.
   (c) Determine what data to collect in order to maximise the probability that the data collected is predictive with respect to the pre-established business objectives. Here, it is vital to include in as much domain specific knowledge as possible from those "on-the-ground". It is also important to account for the principle spatial and temporal scales in the problem and any spatial or temporal inhomogeneities that may be present.
   (d) Determine what part of the data already lies within the organisation and what needs to be obtained externally
   (e) Obtain the data
2. "Build" the Predictability Landscape
   (a) Coarse grain the data appropriately. For example, determine the appropriate "bin" sizes. If the previous Plan to Predict steps are performed successfully a large part of this task should already have been carried out.
   (b) Assess the capabilities of the data analysts, as this will determine what is an appropriate set of statistical tools to use.
   (c) Take a multi-perspective approach (the outcomes of the following steps will normally depend on the particular business objective considered)
      i. Dimensionally reduce the problem via a further coarse graining by determining the important features and feature values using feature detection agents.
      ii. Determine the most appropriate statistical tools for analysing the coarse grained data supported by the feature selection agents. This will depend on whether a profile or a prediction is required, on the data set available and on the nature of the underlying problem - classification, estimation, clustering/segmentation, etc.
      iii. Integrate the information from the different agents
      iv. Make actionable recommendations

We may also add the following wisdom learned from experience in many data mining projects. Some aspects apply more to modellers, others more to managers.

1. Don't be over reliant on statistical software packages - too often they're an excuse for not thinking.
2. Start with simple models then build up Predictability from there. Don't be seduced by sophisticated and subtle techniques, such as neural nets or evolutionary algorithms, unless you know how to use them well.
3. Finding and developing good data mining models is not formulaic but requires many elements and a sufficient amount of time to think.
4. Understand what the data is telling you - there is no substitute for active immersion in the problem you're trying to solve; domain specific knowledge will greatly benefit any data mining project.
5. Finding high levels of Predictability isn't enough. One needs to have a pitch to upper management in terms of concrete, actionable recommendations with a clear business goal.
6. If little to no Predictability is found, be prepared to admit that the data may be "dross" and more predictive data has to be obtained

10 Conclusions

Data mining is the search for Predictability, a concept which can mathematically represent a multitude of different potential business objectives. It can be pictured geometrically as the search for the peaks, valleys and other topographical features of a corresponding Predictability Landscape. In this landscape, Predictability is a stochastic height function. The landscape's topographical features determine how well a particular data mining model will access any Predictability inherent in the data. Predictability is not an inherent feature of a data set, but rather must be sought by first obtaining a data set that is intuitively likely to contain predictive variables - i.e., one must Plan to Predict. Different data mining models, ranging from the simple, such as a linear regression, to the complex, such as a neural network, are more or less suited to model specific aspects of the landscape according to the model bias. As business goals are inevitably multi-objective, there exists a different Predictability Landscape for every objective. As there are no "magic bullet" data mining models valid across various different types of landscape, a multi-perspective approach, where information from various different models is integrated together, is advantageous. This multi-perspective approach can be naturally framed in the anthropomorphic language of "agents". A multi-perspective agent-based approach, compared to more traditional approaches, gives solutions which consider a multitude of models, thus ameliorating weaknesses and reinforcing the advantages of a given method. It offers more accurate and robust predictions, while at the same time allowing for transparent, rich and informative profiles.

In this article, due to space limitations and for concreteness, we have mainly had classification problems in mind, but what we have said about Plan to Predict, the Predictability Landscape and the Multi-perspective approach hold generally, be it time series analysis, text analysis, clustering or any other general area of data mining. Similarly, in terms of explicit statistical models we have limited detailed discussion to a small class while making general comments about other classes. There is an enormous number of potential model types - discriminant analysis, neural nets, decision trees, latent variable models, hidden Markov models, factor analysis and so on. By consulting some of the extensive resources cited in the bibliography and, more importantly, by obtaining practical "hands-on" experience, one may develop an intuition for the relative advantages and disadvantages of these models and what is their most appropriate role in a multi-perspective approach.

Data mining is a difficult area because of the enormous breadth of concepts, tools and techniques that must be mastered in order to do it well. In the current world of ever increasing data streams it is one of great importance, an importance that can only increase in the future.

Acknowledgements

CRS acknowledges support from: CONACyT project 30422-E, a DGAPA Sabatical Fellowship, hospitality and financial support from the Dublin Institute for Advanced Studies and support from the EPSRC (grant number GR/T24616/01). CRS would like to thank Henri Waelbroeck, Susan Talley, Bill Langdon and Rosario Cruz, as well as many other colleagues, for useful conversations.

References

1. E. Arnould, Deep Engagement with Consumer Experience: Listening and Learning with Qualitative Data, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
2. N. Malhotra, Questionnaire Design and Scale Development, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
3. J. Miller, Online Market Research, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
4. S.M. Smith, Advanced Techniques and Technologies in Online Research, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
5. D. Mallett, Sampling and Weighting, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
6. S.M. Smith and G. Albaum, Basic Data Analysis, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
7. D.R. Lehmann, Regression Models, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
8. S. Gupta, Advanced Regression Models, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).

9. S. Cohen, J. Louviere and T. Eagle, Conjoint Models, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
10. P. Bentler and V. Savalei, Structural Equation Models, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
11. S. Sharma and A. Kumar, Cluster Analysis and Factor Analysis, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
12. W. DeSarbo, W. Kamakura and M. Wedel, Latent Structure Regression, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
13. G. Allenby, P. Rossi and R. McCulloch, Hierarchical Bayes Models, In The Handbook of Market Research: Dos and Don'ts, eds. R. Grover and M. Vriens (Sage Publications 2005).
14. M.J.A. Berry and G. Linhoff, Data Mining Techniques for Marketing, Sales and Customer Support, Wiley (1997); *Accessible: descriptive and easily understandable, but discussion of different data mining models is technically weak.*
15. The Handbook of Data Mining, editor Nong Ye, Lawrence Erlaum, NJ (2003); *Advanced: very good set of articles but treatment is dense and technical.*
16. Fayyad, U.M., Piatestsky-Shapiro, G., Smyth, D. and Uthurusamy, R. Advances in Knowledge Discovery and Data Mining, (Cambridge MA AAAI Press/MIT Press 1996). *Advanced: some of the articles may be too specialized to be of use, but some nice general ones too.*
17. S.M. Weiss and N. Indurkhya, Predictive Data Mining: A Practical Guide, (Morgan Kaufmann, San Francisco CA 1998). *Accessible: Nice concise overview of data mining.*
18. David J. Hand, Padhraic Smyth, Heikki Mannila, Principles of Data Mining (MIT Press 2001). *Advanced, but mathematical level is quite accessible. Written by very competant academics.*
19. R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley-Interscience (2000). *Advanced: A classic text; concerned with pattern recognition rather than data mining per se, but many of the important issues and techniques used are the same. Technically sound and comprehensive in its coverage of statistical data mining models. Quite readable too, but relatively weak on the feature selection process.*
20. A. Berson, Stephen Smith, K. Thearling, Building Data Mining Applications for CRM (McGraw Hill 1999) *Accessible: Probably a good introduction for non-technicals with an emphasis on Customer Resource Management.*
21. www.statsoft.com/textbook/stathome.html - *Online textbook at quite accessible level. Contains treatments of all the standard statistical models as well as more advanced techniques.*
22. Andrew Moore's Statistical Data Mining Tutorials at Carnegie Mellon http://www-2.cs.cmu.edu/ awm/tutorials/
23. http://www.scd.ucar.edu/hps/GROUPS/dm/dm.html *Good comprehensive source for datamining resources.*
24. For those with interests in high level computing aspects of data mining, go to the National Center for Data Mining http://www.ncdm.uic.edu
25. Data Mining and Knowledge Discovery - scholarly journal published by Kluwer http://www.kluweronline.com/issn/1384-5810
26. D.L. Kreher and D.R. Stinson, Combinatorial Algorithms: Generation, Enumeration and Search (CRC Press 1998). *Advanced*
27. D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley. *Accessible*

28. T. Back, Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms (OUP 1996). *Advanced*
29. F. Glover and M. Laguna, Tabu Search, (Springer 1998).
30. P.J.M. Van Laarhoven and E.H.L. Aarts, Simulated Annealing: Theory and Applications, (D. Riedel, Dordrecht 1987). *Advanced*
31. J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection (MIT Press 1992). *Advanced*
32. W.B. Langdon and R. Poli, Foundations of Genetic Programming (Springer 2002). *Advanced*
33. P. Domingos and M. Pazzani, *Beyond independence: Conditions for the optimality of the simple Bayesian classifier*, in *Proceedings of the 13th International Conference on Machine Learning*, 105-112 (Morgan Kaufmann 1996).
34. For a list of resources on agents see, for instance the website - http://www.aaai.org/AITopics/html/agents.html
35. P. van der Putten and M. van Someren (eds). *CoIL Challenge 2000: The Insurance Company Case*. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.
36. C. Elkan. Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000. In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01), pp. 426-433