Chapter 1
Introduction

Every day people are faced with questions such as "What route should I take to work today?" "Should I switch to a different cell phone carrier?" "How should I invest my money?" or "Will I get cancer?" These questions indicate our desire to know future events, and we earnestly want to make the best decisions towards that future.

We usually make decisions based on information. In some cases we have tangible, objective data, such as the morning traffic or weather report. Other times we use intuition and experience like "I should avoid the bridge this morning because it usually gets bogged down when it snows" or "I should have a PSA test because my father got prostate cancer." In either case, we are predicting future events given the information and experience we currently have, and we are making decisions based on those predictions.

As information has become more readily available via the internet and media, our desire to use this information to help us make decisions has intensified. And while the human brain can consciously and subconsciously assemble a vast amount of data, it cannot process the even greater amount of easily obtainable, relevant information for the problem at hand. To aid in our decision-making processes, we now turn to tools like Google to filter billions of web pages to find the most appropriate information for our queries, WebMD to diagnose our illnesses based on our symptoms, and E*TRADE to screen thousands of stocks and identify the best investments for our portfolios.

These sites, as well as many others, use tools that take our current information, sift through data looking for patterns that are relevant to our problem, and return answers. The process of developing these kinds of tools has evolved throughout a number of fields such as chemistry, computer science, physics, and statistics and has been called "machine learning," "artificial intelligence," "pattern recognition," "data mining," "predictive analytics," and "knowledge discovery." While each field approaches the problem using different perspectives and tool sets, the ultimate objective is the same: *to make an accurate prediction*. For this book, we will pool these terms into the commonly used phrase *predictive modeling*.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3_1,

Geisser (1993) defines predictive modeling as "the process by which a model is created or chosen to try to best predict the probability of an outcome." We tweak this definition slightly:

> **Predictive modeling**: the process of developing a mathematical tool or model that generates an accurate prediction

Steve Levy of *Wired* magazine recently wrote of the increasing presence of predictive models (Levy 2010), "Examples [of artificial intelligence] can be found everywhere: The Google global machine uses AI to interpret cryptic human queries. Credit card companies use it to track fraud. Netflix uses it to recommend movies to subscribers. And the financial system uses it to handle billions of trades (with only the occasional meltdown)." Examples of the types of questions one would like to predict are:

- How many copies will this book sell?
- Will this customer move their business to a different company?
- How much will my house sell for in the current market?
- Does a patient have a specific disease?
- Based on past choices, which movies will interest this viewer?
- Should I sell this stock?
- Which people should we match in our online dating service?
- Is an e-mail spam?
- Will this patient respond to this therapy?

Insurance companies, as another example, must predict the risks of potential auto, health, and life policy holders. This information is then used to determine if an individual will receive a policy, and if so, at what premium. Like insurance companies, governments also seek to predict risks, but for the purpose of protecting their citizens. Recent examples of governmental predictive models include biometric models for identifying terror suspects, models of fraud detection (Westphal 2008), and models of unrest and turmoil (Shachtman 2011). Even a trip to the grocery store or gas station [everyday places where our purchase information is collected and analyzed in an attempt to understand who we are and what we want (Duhigg 2012)] brings us into the predictive modeling world, and we're often not even aware that we've entered it. Predictive models now *permeate our existence*.

While predictive models guide us towards more satisfying products, better medical treatments, and more profitable investments, they regularly generate inaccurate predictions and provide the wrong answers. For example, most of us have not received an important e-mail due to a predictive model (a.k.a. e-mail filter) that incorrectly identified the message as spam. Similarly, predictive models (a.k.a. medical diagnostic models) misdiagnose diseases, and predictive models (a.k.a. financial algorithms) erroneously buy and sell stocks predicting profits when, in reality, finding losses. This final example of predictive models gone wrong affected many investors in 2010. Those who follow the stock market are likely familiar with the "flash crash" on May 6, 2010,

in which the market rapidly lost more than 600 points, then immediately regained those points. After months of investigation, the Commodity Futures Trading Commission and the Securities and Exchange Commission identified an erroneous algorithmic model as the cause of the crash (U.S. Commodity Futures Trading Commission and U.S. Securities & Exchange Commission 2010).

Stemming in part from the flash crash and other failures of predictive models, Rodriguez (2011) writes, "Predictive modeling, the process by which a model is created or chosen to try to best predict the probability of an outcome has lost credibility as a forecasting tool." He hypothesizes that predictive models regularly fail because they do not account for complex variables such as human behavior. Indeed, our abilities to predict or make decisions are constrained by our present and past knowledge and are affected by factors that we have not considered. These realities are limits of any model, yet these realities should not prevent us from seeking to improve our process and build better models.

There are a number of common reasons why predictive models fail, and we address each of these in subsequent chapters. The common culprits include (1) inadequate pre-processing of the data, (2) inadequate model validation, (3) unjustified extrapolation (e.g., application of the model to data that reside in a space which the model has never seen), or, most importantly, (4) over-fitting the model to the existing data. Furthermore, predictive modelers often only explore relatively few models when searching for predictive relationships. This is usually due to either modelers' preference for, knowledge of, or expertise in, only a few models or the lack of available software that would enable them to explore a wide range of techniques.

This book endeavors to help predictive modelers produce reliable, trustworthy models by providing a step-by-step guide to the model building process and to provide intuitive knowledge of a wide range of common models. The objectives of this book are to provide:

- Foundational principles for building predictive models
- Intuitive explanations of many commonly used predictive modeling methods for both classification and regression problems
- Principles and steps for validating a predictive model
- Computer code to perform the necessary foundational work to build and validate predictive models

To illustrate these principles and methods, we will use a diverse set of real-world examples ranging from finance to pharmaceutical which we describe in detail in Sect. 1.4. But before describing the data, we first explore a reality that confronts predictive modeling techniques: the trade-off between prediction and interpretation.

1.1 Prediction Versus Interpretation

For the examples listed above, historical data likely exist that can be used to create a mathematical tool to predict future, unseen cases. Furthermore, the foremost objective of these examples is not to understand why something will (or will not) occur. Instead, we are primarily interested in accurately projecting the chances that something will (or will not) happen. Notice that the focus of this type of modeling is to optimize prediction accuracy. For example, we don't really care why an e-mail filter thinks a message is spam. Rather, we only care that the filter accurately trashes spam and allows messages we care about to pass through to our mailbox. As another example, if I am selling a house, my primary interest is not how a web site (such as zillow.com) estimated its value. Instead, I am keenly interested that zillow.com has correctly priced the home. An undervaluation will yield lower bids and a lower sale price; alternatively, an overvaluation may drive away potential buyers.

The tension between prediction and interpretation is also present in the medical field. For example, consider the process that a cancer patient and physician encounter when contemplating changing treatment therapies. There are many factors for the physician and patient to consider such as dosing schedule, potential side effects, and survival rates. However, if enough patients have taken the alternative therapy, then data could be collected on these patients related to their disease, treatment history, and demographics. Also, laboratory tests could be collected related to patients' genetic background or other biological data (e.g., protein measurements). Given their outcome, a predictive model could be created to predict the response to the alternative therapy based on these data. The critical question for the doctor and patient is a prediction of *how* the patient will react to a change in therapy. Above all, this prediction needs to be accurate. If a model is created to make this prediction, it should not be constrained by the requirement of interpretability. A strong argument could be made that this would be unethical. As long as the model can be appropriately validated, it should not matter whether it is a black box or a simple, interpretable model.

While the primary interest of predictive modeling is to generate accurate predictions, a secondary interest may be to interpret the model and understand why it works. The unfortunate reality is that as we push towards higher accuracy, models become more complex and their interpretability becomes more difficult. This is almost always the trade-off we make when prediction accuracy is the primary goal.

1.2 Key Ingredients of Predictive Models

The colloquial examples thus far have illustrated that data, in fact very large data sets, can now be easily generated in an attempt to answer almost any type of research question. Furthermore, free or relatively inexpensive model building software such as JMP, WEKA, and many packages in R, as well as powerful personal computers, make it relatively easy for anyone with some computing knowledge to begin to develop predictive models. But as Rodriguez (2011) accurately points out, the credibility of model building has weakened especially as the window to data access and analysis tools has widened.

As we will see throughout this text, if a predictive signal exists in a set of data, many models will find some degree of that signal regardless of the technique or care placed in developing the model. Naïve model application can therefore be effective to an extent; as the saying goes, "even a blind squirrel finds a nut." But the best, most predictive models are fundamentally influenced by a modeler with expert knowledge and context of the problem. This expert knowledge should first be applied in obtaining *relevant* data for the desired research objectives. While vast databases of information can be used as substrate for constructing predictions, irrelevant information can drive down predictive performance of many models. Subject-specific knowledge can help separate potentially meaningful information from irrelevant information, eliminating detrimental noise and strengthening the underlying signal. Undesirable, confounding signal may also exist in the data and may not be able to be identified without expert knowledge. As an extreme example of misleading signal and the need for an expert understanding of the problem, consider the U.S. Food and Drug Administration's Adverse Event Reporting System database which provides information on millions of reported occurrences of drugs and their reported side effects. Obvious biases abound in this collection of data; for example, a search on a drug for treating nausea may reflect that a large proportion of the patients using the treatment had leukemia. An uninformed analysis may identify leukemia as a potential side effect of the drug. The more likely explanation is that the subjects were taking the nausea medication to mitigate the side effects of the cancer therapy. This may be intuitively obvious, but clearly the availability of large quantities of records is not a protection against an uninformed use of the data.

Ayres (2007) extensively studies the interplay between expert opinion and empirical, data-driven models makes two important observations bolstering the need for problem-specific knowledge. Firstly,

> "In the end, [predictive modeling] is not a substitute for intuition, but rather a complement"

Simply put, neither data-driven models nor the expert relying solely on intuition will do better than a combination of the two. Secondly,

> "Traditional experts make better decisions when they are provided with the results of statistical prediction. Those who cling to the authority of traditional

> experts tend to embrace the idea of combining the two forms of 'knowledge' by giving the experts 'statistical support' . . . Humans usually make better predictions when they are provided with the results of statistical prediction."

In some cases, such as spam detection, it may be acceptable to let computers do most of the thinking. When the consequences are more serious, such as predicting patient response, a combined approach often leads to better results.

To summarize, the foundation of an effective predictive model is laid with *intuition* and *deep knowledge of the problem context*, which are entirely vital for driving decisions about model development. That process begins with *relevant* data, another key ingredient. The third ingredient is a *versatile* computational toolbox which includes techniques for data pre-processing and visualization as well as a suite of modeling tools for handling a range of possible scenarios such as those that are described in Table 1.1.

1.3 Terminology

As previously noted, "predictive modeling" is one of the many names that refers to the process of uncovering relationships within data for predicting some desired outcome. Since many scientific domains have contributed to this field, there are synonyms for different entities:

- The terms *sample*, *data point*, *observation*, or *instance* refer to a single, independent unit of data, such as a customer, patient, or compound. The term *sample* can also refer to a subset of data points, such as the training set sample. The text will clarify the appropriate context when this term is used.
- The *training set* consists of the data used to develop models while the *test* or *validation* sets are used solely for evaluating the performance of a final set of candidate models.
- The *predictors*, *independent variables*, *attributes*, or *descriptors* are the data used as input for the prediction equation.
- *Outcome*, *dependent variable*, *target*, *class*, or *response* refer to the outcome event or quantity that is being predicted.
- *Continuous* data have natural, numeric scales. Blood pressure, the cost of an item, or the number of bathrooms are all continuous. In the last case, the counts cannot be a fractional number, but is still treated as continuous data.
- *Categorical* data, otherwise known as *nominal*, *attribute*, or *discrete* data, take on specific values that have no scale. Credit status ("good" or "bad") or color ("red," "blue," etc.) are examples of these data.
- *Model building*, *model training*, and *parameter estimation* all refer to the process of using data to determine values of model equations.

1.4 Example Data Sets and Typical Data Scenarios

In later chapters, case studies are used to illustrate techniques. Before proceeding, it may be instructive to briefly explore a few examples of predictive modeling problems and the types of data used to solve them. The focus here is on the diversity of the problems as well as the characteristics of the collected data. Several example data sets originate from machine learning competitions, which provide real-world problems with an (often monetary) incentive for providing the best solution. Such competitions have a long history in predictive modeling and have greatly stimulated the field.

Music Genre

This data set was published as a contest data set on the TunedIT web site (http://tunedit.org/challenge/music-retrieval/genres). In this competition, the objective was to develop a predictive model for classifying music into six categories. In total, there were 12,495 music samples for which 191 characteristics were determined. The response categories were not balanced (Fig. 1.1), with the smallest segment coming from the heavy metal category (7 %) and the largest coming from the classical category (28 %). All predictors were continuous; many were highly correlated and the predictors spanned different scales of measurement. This data collection was created using 60 performers from which 15–20 pieces of music were selected for each performer. Then 20 segments of each piece were parameterized in order to create the final data set. Hence, the samples are inherently not independent of each other.

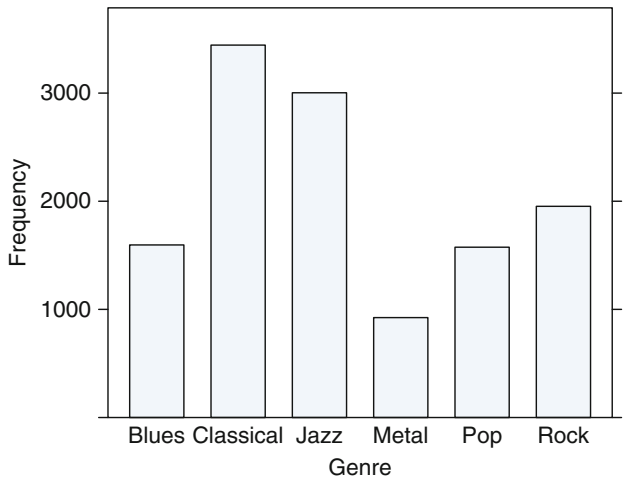

Fig. 1.1: The frequency distribution of genres in the music data

Grant Applications

This data set was also published for a competition on the Kaggle web site (http://www.kaggle.com). For this competition, the objective was to develop a predictive model for the probability of success of a grant application. The historical database consisted of 8,707 University of Melbourne grant applications from 2009 and 2010 with 249 predictors. Grant status (either "unsuccessful" or "successful") was the response and was fairly balanced (46 % successful). The web site notes that current Australian grant success rates are less than 25 %. Hence the historical database rates are not representative of Australian rates. Predictors include measurements and categories such as Sponsor ID, Grant Category, Grant Value Range, Research Field, and Department and were continuous, count, and categorical. Another notable characteristic of this data set is that many predictor values were missing (83 %). Furthermore, the samples were not independent since the same grant writers occurred multiple times throughout the data. These data are used throughout the text to demonstrate different classification modeling techniques.

We will use these data extensively throughout Chaps. 12 through 15, and a more detailed explanation and summary of the data can be found in Sect. 12.1.

Hepatic Injury

A data set from the pharmaceutical industry was used to develop a model for predicting compounds' probability of causing hepatic injury (i.e., liver damage). This data set consisted of 281 unique compounds; 376 predictors were measured or computed for each. The response was categorical (either "does not cause injury," "mild injury," or "severe injury") and was highly unbalanced (Fig. 1.2). This variety of response often occurs in pharmaceutical data because companies steer away from creating molecules that have undesirable safety characteristics. Therefore, well-behaved molecules often greatly outnumber undesirable molecules. The predictors consisted of measurements from 184 biological screens and 192 chemical feature predictors. The biological predictors represent activity for each screen and take values between 0 and 10 with a mode of 4. The chemical feature predictors represent counts of important substructures as well as measures of physical properties that are thought to be associated with hepatic injury. A more extensive description of these types of predictors is given in Chap. 5.

Permeability

This pharmaceutical data set was used to develop a model for predicting compounds' permeability. In short, permeability is the measure of a molecule's ability to cross a membrane. The body, for example, has notable membranes between the body and brain, known as the blood–brain barrier, and between

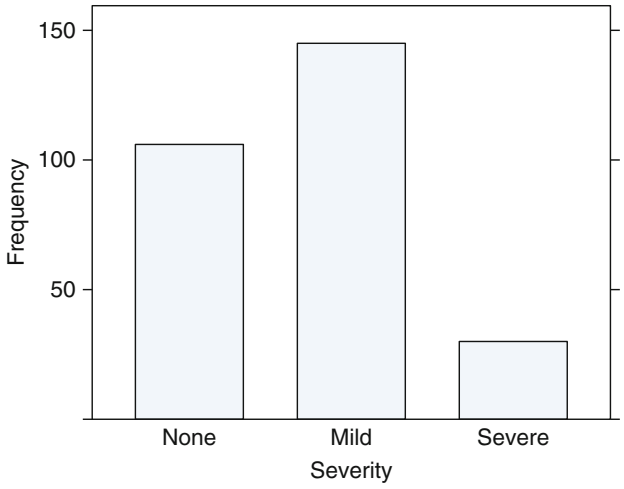

Fig. 1.2: Distribution of hepatic injury type

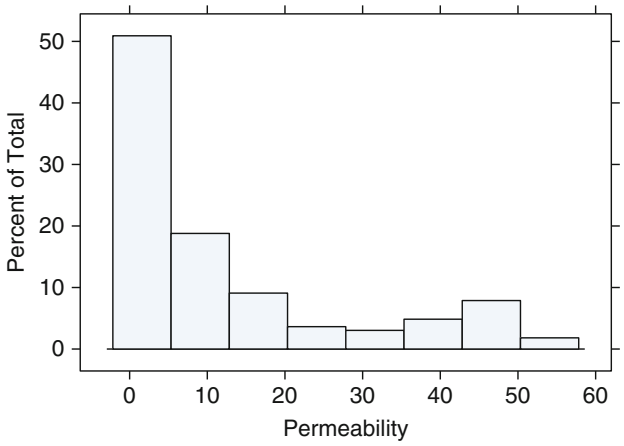

Fig. 1.3: Distribution of permeability values

the gut and body in the intestines. These membranes help the body guard critical regions from receiving undesirable or detrimental substances. For an orally taken drug to be effective in the brain, it first must pass through the intestinal wall and then must pass through the blood–brain barrier in order to be present for the desired neurological target. Therefore, a compound's ability to permeate relevant biological membranes is critically important to understand early in the drug discovery process. Compounds that appear to be effective for a particular disease in research screening experiments but appear to be poorly permeable may need to be altered in order to improve permeability and thus the compound's ability to reach the desired target. Identifying permeability problems can help guide chemists towards better molecules.

Permeability assays such as PAMPA and Caco-2 have been developed to help measure compounds' permeability (Kansy et al. 1998). These screens are effective at quantifying a compound's permeability, but the assay is expensive labor intensive. Given a sufficient number of compounds that have been screened, we could develop a predictive model for permeability in an attempt to potentially reduce the need for the assay. In this project there were 165 unique compounds; 1,107 molecular fingerprints were determined for each. A molecular fingerprint is a binary sequence of numbers that represents the presence or absence of a specific molecular substructure. The response is highly skewed (Fig. 1.3), the predictors are sparse (15.5 % are present), and many predictors are strongly associated.

Chemical Manufacturing Process

This data set contains information about a chemical manufacturing process, in which the goal is to understand the relationship between the process and the resulting final product yield. Raw material in this process is put through a sequence of 27 steps to make the final pharmaceutical product. The starting material is generated from a biological unit and has a range of quality and characteristics. The objective in this project was to develop a model to predict percent yield of the manufacturing process. The data set consisted of 177 samples of biological material for which 57 characteristics were measured. Of the 57 characteristics, there were 12 measurements of the biological starting material and 45 measurements of the manufacturing process. The process variables included measurements such as temperature, drying time, washing time, and concentrations of by-products at various steps. Some of the process measurements can be controlled, while others are observed. Predictors are continuous, count, categorical; some are correlated, and some contain missing values. Samples are not independent because sets of samples come from the same batch of biological starting material.

Fraudulent Financial Statements

Fanning and Cogger (1998) describe a data set used to predict management fraud for publicly traded companies. Using public data sources, such as U.S. Securities and Exchange Commission documents, the authors were able to identify 102 fraudulent financial statements. Given that a small percentage of statements are fraudulent, they chose to sample an equivalent number[1] of non-fraudulent companies, which were sampled to control for important factors (e.g., company size and industry type). Of these data, 150 data points were used to train models and the remaining 54 were used to evaluate them.

[1] This type of sampling is very similar to *case-control studies* in the medical field.

The authors started the analysis with an unidentified number of predictors derived from key areas, such as executive turnover rates, litigation, and debt structure. In the end, they used 20 predictors in their models. Examples include the ratio of accounts receivable to sales, the ratio of inventory to sales, and changes in the gross margins between years. Many of the predictor variables of ratios share common denominators (e.g., the ratio of accounts receivable to sales and the ratio of inventory to sales). Although the actual data points were not published, there is likely to be strong correlations between predictors.

From a modeling perspective, this example is interesting for several reasons. First, because of the large class imbalance, the frequencies of the two classes in the data sets were very different from the population that will be predicted with severe imbalances. This is a common strategy to minimize the consequences of such an imbalance and is sometimes referred to as "down-sampling" the data. Second, the number of possible predictors was large compared to the number of samples. In this situation, the selection of predictors for the models is delicate as there are only a small number of samples for selecting predictors, building models, and evaluating their performance. Later chapters discuss the problem of over-fitting, where trends in the training data are not found in other samples of the population. With a large number of predictors and a small number of data points, there is a risk that a relevant predictor found in this data set will not be reproducible.

Comparisons Between Data Sets

These examples illustrate characteristics that are common to most data sets. First, the response may be continuous or categorical, and for categorical responses there may be more than two categories. For continuous response data, the distribution of the response may be symmetric (e.g., chemical manufacturing) or skewed (e.g., permeability); for categorical response data, the distribution may be balanced (e.g., grant applications) or unbalanced (e.g., music genre, hepatic injury). As we will show in Chap. 4, understanding the distribution of the response is critically necessary for one of the first steps in the predictive modeling process: splitting the data into training and testing sets. Understanding the response distribution will guide the modeler towards better ways of partitioning the data; not understanding response characteristics can lead to computational difficulties for certain kinds of models and to models that have less-than-optimal predictive ability.

The data sets summarized in Table 1.1 also highlight characteristics of predictors that are universal to most data sets. Specifically, the values of predictors may be continuous, count, and/or categorical; they may have missing values and could be on different scales of measurement. Additionally, predictors within a data set may have high correlation or association, thus indicating that the predictor set contains numerically redundant information.

Furthermore, predictors may be sparse, meaning that a majority of samples contain the same information while only a few contain unique information. Like the response, predictors can follow a symmetric or skewed distribution (for continuous predictors) or be balanced or unbalanced (for categorical predictors). Lastly, predictors within a data set may or may not have an underlying relationship with the response.

Different kinds of models handle these types of predictor characteristics in different ways. For example, partial least squares naturally manages correlated predictors but is numerically more stable if the predictors are on similar scales. Recursive partitioning, on the other hand, is unaffected by predictors of different scales but has a less stable partitioning structure when predictors are correlated. As another example of predictor characteristics' impact on models, multiple linear regression cannot handle missing predictor information, but recursive partitioning can be used when predictors contain moderate amounts of missing information. In either of these example scenarios, failure to appropriately adjust the predictors prior to modeling (known as pre-processing) will produce models that have less-than-optimal predictive performance. Assessing predictor characteristics and addressing them through pre-processing is covered in Chap. 3.

Finally, each of these data sets illustrates another fundamental characteristic that must be considered when building a predictive model: the relationship between the number of samples (n) and number of predictors (P). In the case of the music genre data set, the number of samples ($n = 12{,}496$) is much greater than the number of predictors ($P = 191$). All predictive models handle this scenario, but computational time will vary among models and will likely increase as the number of samples and predictors increase. Alternatively, the permeability data set has significantly fewer samples ($n = 165$) than predictors ($P = 1{,}107$). When this occurs, predictive models such as multiple linear regression or linear discriminant analysis cannot be directly used. Yet, some models [e.g., recursive partitioning and K-nearest neighbors (KNNs)] can be used directly under this condition. As we discuss each method in later chapters, we will identify the method's ability to handle data sets where $n < P$. For those that cannot operate under this condition, we will suggest alternative modeling methods or pre-processing steps that will effectively reduce the dimension of the predictor space.

In summary, we must have a detailed understanding of the predictors and the response for any data set prior to attempting to build a model. Lack of understanding can lead to computational difficulties and less than optimal model performance. Furthermore, most data sets will require some degree of pre-processing in order to expand the universe of possible predictive models and to optimize each model's predictive performance.

Table 1.1: A comparison of several characteristics of the example data sets

	Data set					
Data characteristic	Music genre	Grant applications	Hepatic injury	Fraud detection	Permeability	Chemical manufacturing
Dimensions						
# Samples	12,495	8,707	281	204	165	177
# Predictors	191	249	376	20	1,107	57
Response characteristics						
Categorical or continuous	Categorical	Categorical	Categorical	Categorical	Continuous	Continuous
Balanced/symmetric		×		×		×
Unbalanced/skewed	×		×	×		
Independent			×		×	
Predictor Characteristics						
Continuous	×	×	×	×		×
Count	×	×	×			×
Categorical		×	×	×	×	×
Correlated/associated	×	×	×	×	×	×
Different scales	×	×	×	×		×
Missing values		×				×
Sparse					×	