CHAPTER

1

Overview of Predictive Analytics

A small direct response company had developed dozens of programs in cooperation with major brands to sell books and DVDs. These affinity programs were very successful, but required considerable up-front work to develop the creative content and determine which customers, already engaged with the brand, were worth the significant marketing spend to purchase the books or DVDs on subscription. Typically, they first developed test mailings on a moderately sized sample to determine if the expected response rates were high enough to justify a larger program.

One analyst with the company identified a way to help the company become more profitable. What if one could identify the key characteristics of those who responded to the test mailing? Furthermore, what if one could generate a score for these customers and determine what minimum score would result in a high enough response rate to make the campaign profitable? The analyst discovered predictive analytics techniques that could be used for both purposes, finding key customer characteristics and using those characteristics to generate a score that could be used to determine which customers to mail.

Two decades before, the owner of a small company in Virginia had a compelling idea: Improve the accuracy and flexibility of guided munitions using optimal control. The owner and president, Roger Barron, began the process of deriving the complex mathematics behind optimal control using a technique known as variational calculus and hired a graduate student to assist him in the task. Programmers then implemented the mathematics in computer code so

they could simulate thousands of scenarios. For each trajectory, the variational calculus minimized the miss distance while maximizing speed at impact as well as the angle of impact.

The variational calculus algorithm succeeded in identifying the optimal sequence of commands: how much the fins (control surfaces) needed to change the path of the munition to follow the optimal path to the target. The concept worked in simulation in the thousands of optimal trajectories that were run. Moreover, the mathematics worked on several munitions, one of which was the MK82 glide bomb, fitted (in simulation) with an inertial guidance unit to control the fins: an early smart-bomb.

There was a problem, however. The variational calculus was so computationally complex that the small computers on-board could not solve the problem in real time. But what if one could *estimate* the optimal guidance commands at any time during the flight from observable characteristics of the flight? After all, the guidance unit can compute where the bomb is in space, how fast it is going, and the distance of the target that was programmed into the unit when it was launched. If the estimates of the optimum guidance commands were close enough to the actual optimal path, it would be *near optimal* and still succeed. Predictive models were built to do exactly this. The system was called *Optimal Path-to-Go* guidance.

These two programs designed by two different companies seemingly could not be more different. One program knows characteristics of people, such as demographics and their level of engagement with a brand, and tries to predict a human decision. The second program knows locations of a bomb in space and tries to predict the best physical action for it to hit a target.

But they share something in common: They both need to estimate values that are unknown but tremendously useful. For the affinity programs, the models estimate whether or not an individual will respond to a campaign, and for the guidance program, the models estimate the best guidance command. In this sense, these two programs are very similar because they both involve predicting a value or values that are known historically, but are unknown at the time a decision is needed. Not only are these programs related in this sense, but they are far from unique; there are countless decisions businesses and government agencies make every day that can be improved by using historic data as an aid to making decisions or even to automate the decisions themselves.

This book describes the back-story behind how analysts build the predictive models like the ones described in these two programs. There is science behind much of what predictive modelers do, yet there is also plenty of *art*, where no theory can inform us as to the best action, but experience provides principles by which tradeoffs can be made as solutions are found. Without the art, the science would only be able to solve a small subset of problems we face. Without

the science, we would be like a plane without a rudder or a kite without a tail, moving at a rapid pace without any control, unable to achieve our objectives.

What Is Analytics?

Analytics is the process of using computational methods to discover and report influential patterns in data. The goal of analytics is to gain insight and often to affect decisions. Data is necessarily a measure of historic information so, by definition, analytics examines historic data. The term itself rose to prominence in 2005, in large part due to the introduction of Google Analytics. Nevertheless, the ideas behind analytics are not new at all but have been represented by different terms throughout the decades, including *cybernetics*, *data analysis*, *neural networks*, *pattern recognition*, *statistics*, *knowledge discovery*, data *mining*, and now even *data science*.

The rise of analytics in recent years is pragmatic: As organizations collect more data and begin to summarize it, there is a natural progression toward using the data to improve estimates, forecasts, decisions, and ultimately, efficiency.

What Is Predictive Analytics?

Predictive analytics is the process of discovering interesting and meaningful patterns in data. It draws from several related disciplines, some of which have been used to discover patterns in data for more than 100 years, including pattern recognition, statistics, machine learning, artificial intelligence, and data mining. What differentiates predictive analytics from other types of analytics?

First, predictive analytics is data-driven, meaning that algorithms derive key characteristic of the models from the data itself rather than from assumptions made by the analyst. Put another way, data-driven algorithms *induce* models from the data. The induction process can include identification of variables to be included in the model, parameters that define the model, weights or coefficients in the model, or model complexity.

Second, predictive analytics algorithms automate the process of finding the patterns from the data. Powerful induction algorithms not only discover coefficients or weights for the models, but also the very form of the models. Decision trees algorithms, for example, learn which of the candidate inputs best predict a target variable in addition to identifying which values of the variables to use in building predictions. Other algorithms can be modified to perform searches, using exhaustive or greedy searches to find the best set of inputs and model parameters. If the variable helps reduce model error, the variable is included

in the model. Otherwise, if the variable does not help to reduce model error, it is eliminated.

Another automation task available in many software packages and algorithms automates the process of transforming input variables so that they can be used effectively in the predictive models. For example, if there are a hundred variables that are candidate inputs to models that can be or should be transformed to remove skew, you can do this with some predictive analytics software in a single step rather than programming all one hundred transformations one at a time.

Predictive analytics doesn't do anything that any analyst couldn't accomplish with pencil and paper or a spreadsheet if given enough time; the algorithms, while powerful, have no common sense. Consider a supervised learning data set with 50 inputs and a single binary target variable with values 0 and 1. One way to try to identify which of the inputs is most related to the target variable is to plot each variable, one at a time, in a histogram. The target variable can be superimposed on the histogram, as shown in Figure 1-1. With 50 inputs, you need to look at 50 histograms. This is not uncommon for predictive modelers to do.

If the patterns require examining two variables at a time, you can do so with a scatter plot. For 50 variables, there are 1,225 possible scatter plots to examine. A dedicated predictive modeler might actually do this, although it will take some time. However, if the patterns require that you examine three variables simultaneously, you would need to examine 19,600 3D scatter plots in order to examine all the possible three-way combinations. Even the most dedicated modelers will be hard-pressed to spend the time needed to examine so many plots.

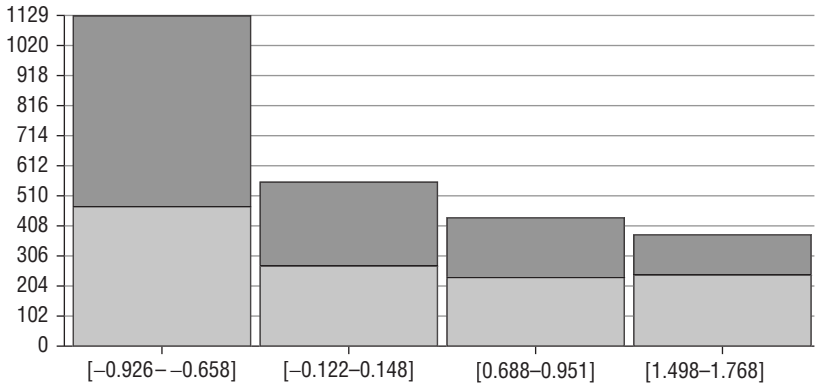

Figure 1-1: Histogram

You need algorithms to sift through all of the potential combinations of inputs in the data—the patterns—and identify which ones are the most interesting. The analyst can then focus on these patterns, undoubtedly a much smaller number of inputs to examine. Of the 19,600 three-way combinations of inputs, it may

be that a predictive model identifies six of the variables as the most significant contributors to accurate models. In addition, of these six variables, the top three are particularly good predictors and much better than any two variables by themselves. Now you have a manageable subset of plots to consider: 63 instead of nearly 20,000. This is one of the most powerful aspects of predictive analytics: identifying which inputs are the most important contributors to patterns in the data.

Supervised vs. Unsupervised Learning

Algorithms for predictive modeling are often divided into two groups: supervised learning methods and unsupervised learning methods. In supervised learning models, the *supervisor* is the target variable, a column in the data representing values to predict from other columns in the data. The target variable is chosen to represent the answer to a question the organization would like to answer or a value unknown at the time the model is used that would help in decisions. Sometimes supervised learning is also called *predictive modeling*. The primary predictive modeling algorithms are *classification* for categorical target variables or *regression* for continuous target variables.

Examples of target variables include whether a customer purchased a product, the amount of a purchase, if a transaction was fraudulent, if a customer stated they enjoyed a movie, how many days will transpire before the next gift a donor will make, if a loan defaulted, and if a product failed. Records without a value for the target variable cannot be used in building predictive models.

Unsupervised learning, sometimes called *descriptive modeling*, has no target variable. The inputs are analyzed and grouped or clustered based on the proximity of input values to one another. Each group or cluster is given a label to indicate which group a record belongs to. In some applications, such as in customer analytics, unsupervised learning is just called segmentation because of the function of the models (segmenting customers into groups).

The key to supervised learning is that the inputs to the model are known but there are circumstances where the target variable is unobserved or unknown. The most common reason for this is a target variable that is an event, decision, or other behavior that takes place at a time future to the observed inputs to the model. Response models, cross-sell, and up-sell models work this way: Given what is known now about a customer, can you predict if they will purchase a particular product in the future?

Some definitions of *predictive analytics* emphasize the function of algorithms as forecasting or predicting *future* events or behavior. While this is often the case, it certainly isn't always the case. The target variable could represent an unobserved variable like a missing value. If a taxpayer didn't file a return in a prior year, predictive models can predict that missing value from other examples of tax returns where the values are known.

Parametric vs. Non-Parametric Models

Algorithms for predictive analytics include both parametric and non-parametric algorithms. Parametric algorithms (or models) assume known distributions in the data. Many parametric algorithms and statistical tests, although not all, assume normal distributions and find linear relationships in the data. Machine learning algorithms typically do not assume distributions and therefore are considered non-parametric or distribution-free models.

The advantage of parametric models is that if the distributions are known, extensive properties of the data are also known and therefore algorithms can be proven to have very specific properties related to errors, convergence, and certainty of learned coefficients. Because of the assumptions, however, the analyst often spends considerable time transforming the data so that these advantages can be realized.

Non-parametric models are far more flexible because they do not have underlying assumptions about the distribution of the data, saving the analyst considerable time in preparing data. However, far less is known about the data *a priori*, and therefore non-parametric algorithms are typically iterative, without any guarantee that the best or optimal solution has been found.

Business Intelligence

Business intelligence is a vast field of study that is the subject of entire books; this treatment is brief and intended to summarize the primary characteristics of business intelligence as they relate to predictive analytics. The output of many business intelligence analyses are reports or dashboards that summarize interesting characteristics of the data, often described as Key Performance Indicators (KPIs). The KPI reports are user-driven, determined by an analyst or decision-maker to represent a key descriptor to be used by the business. These reports can contain simple summaries or very complex, multidimensional measures. Interestingly, KPI is almost never used to describe measures of interest in predictive analytics software and conferences.

Typical business intelligence output is a report to be used by analysts and decision-makers. The following are typical questions that might be answered by business intelligence for fraud detection and customer analytics:

Fraud Detection

- How many cases were investigated last month?
- What was the success rate in collecting debts?
- How much revenue was recovered through collections?
- What was the ROI for the various collection avenues: letters, calls, agents?

- What was the close rate of cases in the past month? Past quarter? Past year?
- For debts that were closed out, how many days did it take on average to close out debts?
- For debts that were closed out, how many contacts with the debtor did it take to close out debt?

Customer Analytics

- What were the e-mail open, click-through, and response rates?
- Which regions/states/ZIPs had the highest response rates?
- Which products had the highest/lowest click-through rates?
- How many repeat purchasers were there last month?
- How many new subscriptions to the loyalty program were there?
- What is the average spend of those who belong to the loyalty program? Those who aren't a part of the loyalty program? Is this a significant difference?
- How many visits to the store/website did a person have?

These questions describe characteristics of the unit of analysis: a customer, a transaction, a product, a day, or even a ZIP code. Descriptions of the unit of analysis are contained in the columns of the data: the attributes. For fraud detection, the unit of analysis is sometimes a debt to be collected, or more generally a case. For customer analytics, the unit of analysis is frequently a customer but could be a visit (a single customer could visit many times and therefore will appear in the data many times).

Note that often these questions compare directly one attribute of interest with an outcome of interest. These questions were developed by a domain expert (whether an analyst, program manager, or other subject matter expert) as a way to describe interesting relationships in the data relevant to the company. In other words, these measures are user-driven.

Are these KPIs and reports actionable decisions in and of themselves? The answer is no, although they can be with small modifications. In the form of the report, you know what happened and can even identify why it happened in some cases. It isn't a great leap, however, to take reports and turn them into predictions. For example, a report that summarizes the response rates for each ZIP code can then use ZIP as a predictor of response rate.

If you consider the reports related to a target variable such as response rate, the equivalent machine learning approach is building a *decision stump*, a single condition rule that predicts the outcome. But this is a very simple way of approaching prediction.

Predictive Analytics vs. Business Intelligence

What if you reconstruct the two lists of questions in a different way, one that is focused more directly on decisions? From a predictive analytics perspective, you may find these questions are the ones asked.

Fraud Detection

- What is the likelihood that the transaction is fraudulent?
- What is the likelihood the invoice is fraudulent or warrants further investigation?
- Which characteristics of the transaction are most related to or most predictive of fraud (single characteristics and interactions)?
- What is the expected amount of fraud?
- What is the likelihood that a tax return is non-compliant?
- Which line items on a tax return contribute the most to the fraud score?
- Historically, which demographic and historic purchase patterns were most related to fraud?

Customer Analytics for Predictive Analytics

- What is the likelihood an e-mail will be opened?
- What is the likelihood a customer will click-through a link in an e-mail?
- Which product is a customer most likely to purchase if given the choice?
- How many e-mails should the customer receive to maximize the likelihood of a purchase?
- What is the best product to up-sell to the customer after they purchase a product?
- What is the visit volume expected on the website next week?
- What is the likelihood a product will sell out if it is put on sale?
- What is the estimated customer lifetime value (CLV) of each customer?

Notice the differences in the kinds of questions predictive analytics asks compared to business intelligence. The word "likelihood" appears often, meaning we are computing a probability that the pattern exists for a unit of analysis. In customer analytics, this could mean computing a probability that a customer is likely to purchase a product.

Implicit in the wording is that the measures require an examination of the groups of records comprising the unit of analysis. If the likelihood an individual customer will purchase a product is one percent, this means that for every 100 customers with the same pattern of measured attributes for this customer,

one customer purchased the product in the historic data used to compute the likelihood. The comparable measure in the business intelligence lists would be described as a *rate* or a *percentage*; what is the response rate of customers with a particular purchase pattern.

The difference between the business intelligence and predictive analytics measures is that the business intelligence variables identified in the questions were, as already described, user driven. In the predictive analytics approach, the predictive modeling algorithms considered many patterns, sometimes all possible patterns, and determined which ones were most predictive of the measure of interest (likelihood). The discovery of the patterns is data driven.

This is also why many of the questions begin with the word "which." Asking *which* line items on a tax return are most related to noncompliance requires comparisons of the line items as they relate to noncompliance.

Do Predictive Models Just State the Obvious?

Often when presenting models to decision-makers, modelers may hear a familiar refrain: "I didn't need a model to tell me that!" But predictive models do more than just identify attributes that are related to a target variable. They identify the *best way* to predict the target. Of all the possible alternatives, all of the attributes that could predict the target and all of the interactions between the attributes, which combinations do the best job? The decision-maker may have been able to guess (hypothesize) that length or residence is a good attribute to predict a responder to a Medicare customer acquisition campaign, but that same person may not have known that the number of contacts is even more predictive, especially when the prospect has been mailed two to six times. Predictive models identify not only which variables are predictive, but how well they predict the target. Moreover, they also reveal which combinations are not just predictive of the target, but *how well* the combinations predict the target and how much better they predict than individual attributes do on their own.

Similarities between Business Intelligence and Predictive Analytics

Often, descriptions of the differences between business intelligence and predictive analytics stress that business intelligence is retrospective analysis, looking back into the past, whereas predictive analytics or prospective analysis predict future behavior. The "predicting the future" label is applied often to predictive analytics in general and the very questions described already imply this is the case. Questions such as "What is the likelihood a customer will purchase . . ." are forecasting future behavior.

Figure 1-2 shows a timeline relating data used to build predictive models or business intelligence reports. The vertical line in the middle is the time the

model is being built (today). The data used to build the models is always to the left: historic data. When predictive models are built to predict a "future" event, the data selected to build the predictive models is rolled back to a time prior to the date the future event is known.

For example, if you are building models to predict whether a customer will respond to an e-mail campaign, you begin with the date the campaign cured (when all the responses have come in) to identify everyone who responded. This is the date for the label "target variable computed based on this date" in the figure. The attributes used as inputs must be known prior to the date of the mailing itself, so these values are collected to the left of the target variable collection date. In other words, the data is set up with all the modeling data in the past, but the target variable is still future to the date the attributes are collected in the timeline of the data used for modeling.

However, it's important to be clear that both business intelligence and predictive analytics analyses are built from the same data, and the data is historic in both cases. The assumption is that future behavior to the right of the vertical line in Figure 1-2 will be consistent with past behavior. If a predictive model identifies patterns in the past that predicted (in the past) that a customer would purchase a product, you assume this relationship will continue to be present in the future.

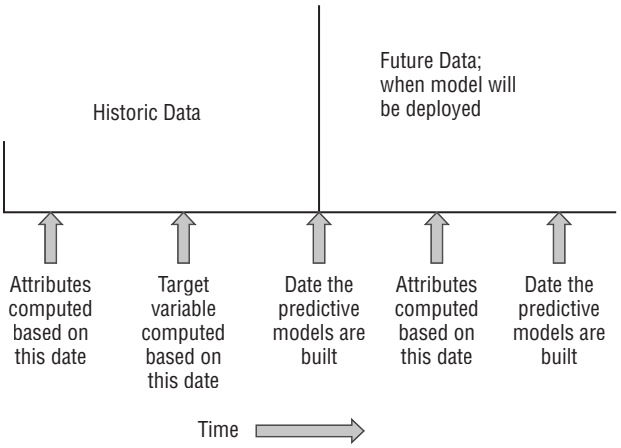

Figure 1-2: Timeline for building predictive models

Predictive Analytics vs. Statistics

Predictive analytics and statistics have considerable overlap, with some statisticians arguing that predictive analytics is, at its core, an extension of statistics. Predictive modelers, for their part, often use algorithms and tests common in statistics as a part of their regular suite of techniques, sometimes without

applying the diagnostics most statisticians would apply to ensure the models are built properly.

Since predictive analytics draws heavily from statistics, the field has taken to heart the amusing quote from statistician and creator of the bootstrap, Brad Efron: "Those who ignore Statistics are condemned to reinvent it." Nevertheless, there are significant differences between typical approaches of the two fields. Table 1-1 provides a short list of items that differ between the fields. Statistics is driven by theory in a way that predictive analytics is not, where many algorithms are drawn from other fields such as machine learning and artificial intelligence that have no provable optimum solution.

But perhaps the most fundamental difference between the fields is summarized in the last row of the table: For statistics, the model is king, whereas for predictive analytics, data is king.

Table 1-1: Statistics vs. Predictive Analytics

STATISTICS	PREDICTIVE ANALYTICS
Models based on theory: There is an optimum.	Models often based on non-parametric algorithms; no guaranteed optimum
Models typically linear.	Models typically nonlinear
Data typically smaller; algorithms often geared toward accuracy with small data	Scales to big data; algorithms not as efficient or stable for small data
The model is king.	Data is king.

Statistics and Analytics

In spite of the similarities between statistics and analytics, there is a difference in mindset that results in differences in how analyses are conducted. Statistics is often used to perform confirmatory analysis where a hypothesis about a relationship between inputs and an output is made, and the purpose of the analysis is to confirm or deny the relationship and quantify the degree of that confirmation or denial. Many analyses are highly structured, such as determining if a drug is effective in reducing the incidence of a particular disease.

Controls are essential to ensure that bias is not introduced into the model, thus misleading the analyst's interpretation of the model. Coefficients of models are critically important in understanding what the data are saying, and therefore great care is taken to transform the model inputs and outputs so they comply with assumptions of the modeling algorithms. If the study is predicting the effect of caloric intake, smoking, age, height, amount of exercise, and metabolism on an individual's weight, and one is to trust the relative contribution of each factor on an individual's weight, it is important to remove any bias due to the data itself so that the conclusions reflect the intent of the model. Bias in

the data could result in misleading the analyst that the inputs to the model have more or less influence that they actually have, simply because of numeric problems in the data.

Residuals are also carefully examined to identify departure from a Normal distribution, although the requirement of normality lessens as the size of the data increases. If residuals are not random with constant variance, the statistician will modify the inputs and outputs until these problems are corrected.

Predictive Analytics and Statistics Contrasted

Predictive modelers, on the other hand, often show little concern for final parameters in the models except in very general terms. The key is often the predictive accuracy of the model and therefore the ability of the model to make and influence decisions. In contrast to the structured problem being solved through confirmatory analysis using statistics, predictive analytics often attempts to solve less structured business problems using data that was not even collected for the purpose of building models; it just happened to be around. Controls are often not in place in the data and therefore causality, very difficult to uncover even in structured problems, becomes exceedingly difficult to identify. Consider, for example, how you would go about identifying which marketing campaign to apply to a current customer for a digital retailer. This customer could receive content from any one of ten programs the e-mail marketing group has identified. The modeling data includes customers, their demographics, their prior behavior on the website and with e-mail they had received in the past, and their reaction to sample content from one of the ten programs. The reaction could be that they ignored the e-mail, opened it, clicked through the link, and ultimately purchased the product promoted in the e-mail. Predictive models can certainly be built to identify the best program of the ten to put into the e-mail based on a customer's behavior and demographics.

However, this is far from a controlled study. While this program is going on, each customer continues to interact with the website, seeing other promotions. The customer may have seen other display ads or conducted Google searches further influencing his or her behavior. The purpose of this kind of model cannot be to uncover fully why the customer behaves in a particular way because there are far too many unobserved, confounding influences. But that doesn't mean the model isn't useful.

Predictive modelers frequently approach problems in this more unstructured, even casual manner. The data, in whatever form it is found, drives the models. This isn't a problem as long as the data continues to be collected in a manner consistent with the data as it was used in the models; consistency in the data will increase the likelihood that there will be consistency in the model's predictions, and therefore how well the model affects decisions.

Predictive Analytics vs. Data Mining

Predictive analytics has much in common with its immediate predecessor, data mining; the algorithms and approaches are generally the same. Data mining has a history of applications in a wide variety of fields, including finance, engineering, manufacturing, biotechnology, customer relationship management, and marketing. I have treated the two fields as generally synonymous since "predictive analytics" became a popular term.

This general overlap between the two fields is further emphasized by how software vendors brand their products, using both data mining and predictive analytics (some emphasizing one term more than the other).

On the other hand, data mining has been caught up in the specter of privacy concerns, spam, malware, and unscrupulous marketers. In the early 2000s, congressional legislation was introduced several times to curtail specifically any data mining programs in the Department of Defense (DoD). Complaints were even waged against the use of data mining by the NSA, including a letter sent by Senator Russ Feingold to the National Security Agency (NSA) Director in 2006:

> *One element of the NSA's domestic spying program that has gotten too little attention is the government's reportedly widespread use of data mining technology to analyze the communications of ordinary Americans. Today I am calling on the Director of National Intelligence, the Defense Secretary and the Director of the NSA to explain whether and how the government is using data mining technology, and what authority it claims for doing so.*

In an interesting *déjà vu*, in 2013, information about NSA programs that sift through phone records was leaked to the media. As in 2006, concerns about privacy were again raised, but this time the mathematics behind the program, while typically described as data mining in the past, was now often described as predictive analytics.

Graduate programs in analytics often use both data mining and predictive analytics in their descriptions, even if they brand themselves with one or the other.

Who Uses Predictive Analytics?

In the 1990s and early 2000s, the use of advanced analytics, referred to as data mining or computational statistics, was relegated to only the most forward-looking companies with deep pockets. Many organizations were still struggling with collecting data, let alone trying to make sense of it through more advanced techniques.

Today, the use of analytics has moved from a niche group in large organizations to being an instrumental component of most mid- to large-sized organizations.

The analytics often begins with business intelligence and moves into predictive analytics as the data matures and the pressure to produce greater benefit from the data increases. Even small organizations, for-profit and non-profit, benefit from predictive analytics now, often using open source software to drive decisions on a small scale.

Challenges in Using Predictive Analytics

Predictive analytics can generate significant improvements in efficiency, decision-making, and return on investment. But predictive analytics isn't always successful and, in all likelihood, the majority of predictive analytics models are never used operationally.

Some of the most common reasons predictive models don't succeed can be grouped into four categories: obstacles in management, obstacles with data, obstacles with modeling, and obstacles in deployment.

Obstacles in Management

To be useful, predictive models have to be deployed. Often, deployment in of itself requires a significant shift in resources for an organization and therefore the project often needs support from management to make the transition from research and development to operational solution. If program management is not a champion of the predictive modeling project and the resulting models, perfectly good models will go unused due to lack of resources and lack of political will to obtain those resources.

For example, suppose an organization is building a fraud detection model to identify transactions that appear to be suspicious and are in need of further investigation. Furthermore, suppose the organization can identify 1,000 transactions per month that should receive further scrutiny from investigators. Processes have to be put into place to distribute the cases to the investigators, and the fraud detection model has to be sufficiently trusted by the investigators for them to follow through and investigate the cases. If management is not fully supportive of the predictive models, these cases may be delivered but end up dead on arrival.

Obstacles with Data

Predictive models require data in the form of a single table or flat file containing rows and columns: two-dimensional data. If the data is stored in transactional databases, keys need to be identified to join the data from the data sources to form the single view or table. Projects can fail before they even begin if the keys don't exist in the tables needed to build the data.

Even if the data can be joined into a single table, if the primary inputs or outputs are not populated sufficiently or consistently, the data is meaningless. For example, consider a customer acquisition model. Predictive models need examples of customers who were contacted and did *not* respond as well as those who were contacted and *did* respond. If active customers are stored in one table and marketing contacts (leads) in a separate table, several problems can thwart modeling efforts. First, unless customer tables include the campaign they were acquired from, it may be impossible to reconstruct the list of leads in a campaign along with the label that the lead responded or didn't respond to the contact.

Second, if customer data, including demographics (age, income, ZIP), is overwritten to keep it up-to-date, and the demographics at the time they were acquired is not retained, a table containing leads as they appeared at the time of the marketing campaign can never be reconstructed. As a simple example, suppose phone numbers are only obtained after the lead converts and becomes a customer. A great predictor of a lead becoming a customer would then be whether the lead has a phone number; this is leakage of future data unknown at the time of the marketing campaign into the modeling data.

Obstacles with Modeling

Perhaps the biggest obstacle to building predictive models from the analyst's perspective is *overfitting*, meaning that the model is too complex, essentially memorizing the training data. The effect of overfitting is twofold: The model performs poorly on new data and the interpretation of the model is unreliable. If care isn't taken in the experimental design of the predictive models, the extent of model overfit isn't known until the model has already been deployed and begins to fail.

A second obstacle with building predictive models occurs when zealous analysts become too ambitious in the kind of model that can be built with the available data and in the timeframe allotted. If they try to "hit a home run" and can't complete the model in the timeframe, no model will be deployed at all. Often a better strategy is to build simpler models first to ensure a model of some value will be ready for deployment. Models can be augmented and improved later if time allows.

For example, consider a customer retention model for a company with an online presence. A zealous modeler may be able to identify thousands of candidate inputs to the retention model, and in an effort to build the best possible model, may be slowed by the sheer combinatorics involved with data preparation and variable selection prior to and during modeling.

However, from the analyst's experience, he or she may be able to identify 100 variables that have been good predictors historically. While the analyst suspects that a better model could be built with more candidate inputs, the first model can be built from the 100 variables in a much shorter timeframe.

Obstacles in Deployment

Predictive modeling projects can fail because of obstacles in the deployment stage of modeling. The models themselves are typically not very complicated computationally, requiring only dozens, hundreds, thousands, or tens of thousands of multiplies and adds, easily handled by today's servers.

At the most fundamental level, however, the models have to be able to be interrogated by the operational system and to issue predictions consistent with that system. In transactional systems, this typically means the model has to be encoded in a programming language that can be called by the system, such as SQL, C++, Java, or another high-level language. If the model cannot be translated or is translated incorrectly, the model is useless operationally.

Sometimes the obstacle is getting the data into the format needed for deployment. If the modeling data required joining several tables to form the single modeling table, deployment must replicate the same joining steps to build the data the models need for scoring. In some transactional systems with disparate data forming the modeling table, complex joins may not be possible in the timeline needed. For example, consider a model that recommends content to be displayed on a web page. If that model needs data from the historic patterns of browsing behavior for a visitor and the page needs to be rendered in less than one second, all of the data pulls and transformations must meet this timeline.

What Educational Background Is Needed to Become a Predictive Modeler?

Conventional wisdom says that predictive modelers need to have an academic background in statistics, mathematics, computer science, or engineering. A degree in one of these fields is best, but without a degree, at a minimum, one should at least have taken statistics or mathematics courses. Historically, one could not get a degree in predictive analytics, data mining, or machine learning.

This has changed, however, and dozens of universities now offer master's degrees in predictive analytics. Additionally, there are many variants of analytics degrees, including master's degrees in data mining, marketing analytics, business analytics, or machine learning. Some programs even include a practicum so that students can learn to apply textbook science to real-world problems.

One reason the real-world experience is so critical for predictive modeling is that the science has tremendous limitations. Most real-world problems have data problems never encountered in the textbooks. The ways in which data can go wrong are seemingly endless; building the same customer acquisition models even within the same domain requires different approaches to data preparation, missing value imputation, feature creation, and even modeling methods.

However, the *principles* of how one can solve data problems are not endless; the experience of building models for several years will prepare modelers to at least be able to identify when potential problems may arise.

Surveys of top-notch predictive modelers reveal a mixed story, however. While many have a science, statistics, or mathematics background, many do not. Many have backgrounds in social science or humanities. How can this be?

Consider a retail example. The retailer Target was building predictive models to identify likely purchase behavior and to incentivize future behavior with relevant offers. Andrew Pole, a Senior Manager of Media and Database Marketing described how the company went about building systems of predictive models at the Predictive Analytics World Conference in 2010. Pole described the importance of a combination of domain knowledge, knowledge of predictive modeling, and most of all, a forensic mindset in successful modeling of what he calls a "guest portrait."

They developed a model to predict if a female customer was pregnant. They noticed patterns of purchase behavior, what he called "nesting" behavior. For example, women were purchasing cribs on average 90 days before the due date. Pole also observed that some products were purchased at regular intervals prior to a woman's due date. The company also observed that if they were able to acquire these women as purchasers of other products during the time before the birth of their baby, Target was able to increase significantly the customer value; these women would continue to purchase from Target after the baby was born based on their purchase behavior before.

The key descriptive terms are "*observed*" and "*noticed*." This means the models were not built as black boxes. The analysts asked, "does this make sense?" and leveraged insights gained from the patterns found in the data to produce better predictive models. It undoubtedly was iterative; as they "noticed" patterns, they were prompted to consider other patterns they had not explicitly considered before (and maybe had not even occurred to them before). This forensic mindset of analysts, noticing interesting patterns and making connections between those patterns and how the models could be used, is critical to successful modeling. It is rare that predictive models can be fully defined before a project and anticipate all of the most important patterns the model will find. So we shouldn't be surprised that we *will* be surprised, or put another way, we should *expect* to be surprised.

This kind of mindset is not learned in a university program; it is part of the personality of the individual. Good predictive modelers need to have a forensic mindset and intellectual curiosity, whether or not they understand the mathematics enough to derive the equations for linear regression.