Dr. N. SHAKEELA

Guest Lecturer

School of Computer Science, Engineering & Applications

Bharathidasan University

Trichy-24

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.

.

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 prepara- tion, 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 predictivemodeling, and most of all, a forensic mindset in successful modeling**

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.*" , leveraged(manipulated) insights gained from the patterns found in the data to produce better predictive models. It undoubtedly was iterative; as they "noticed" pat- terns, 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..**

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 .

Setting Up the Problem

The most important part of any predictive modeling project is the very beginning when the predictive modeling project is defined. Setting up a predictive modeling project is a very difficult task because the skills needed to do it well are very broad, requiring knowledge of the business domain, databases, or data infrastructure, and predictive modeling algorithms and techniques.

Predictive Analytics Processing Steps: CRISP-DM

The Cross-Industry Standard Process Model for Data Mining (CRISP-DM) describes the data-mining process in six steps. It has been cited as the most-often used process model since its inception in the 1990s.

One advantage of using CRISP-DM is that it describes the most commonly

applied steps in the process and is documented in an 80-page PDF file. The CRISP-DM name itself calls out data mining as the technology, but the same **process model applies to predictive analytics and other related analytics** approaches, including business analytics, statistics, and text mining.

The CRISP-DM audience includes both managers and practitioners. For program managers, Each of the steps can then have its own cost estimates and can be tracked by the manager to ensure the project deliverables and timetables are met. The last step in many of the sub-tasks in CRISP-DM is a report describing what decisions were made and why. In fact, the CRISP-DM document **identifies** 28 potential deliverables for a project.

For **practitioners**, the step-by-step process provides structure for analysis and not only reminds the analyst of the steps that need to be accomplished, but also the need for documentation and reporting throughout the process, which is particularly valuable for new modelers. Even for experienced practitioners, CRISP-DM describes the steps succinctly and logically.

The six steps in the CRISP-DM process are shown in Figure 2-1: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Table 2-1: CRISM-DM Sequence

STAGE	DESCRIPTION
Business Understanding	Define the project.
Data Understanding	Examine the data; identify problems in the data.
Data Preparation	Fix problems in the data; create derived variables.
Modeling	Build predictive or descriptive models.
Evaluation	Assess models; report on the expected effects of models.
Deployment	Plan for use of models.

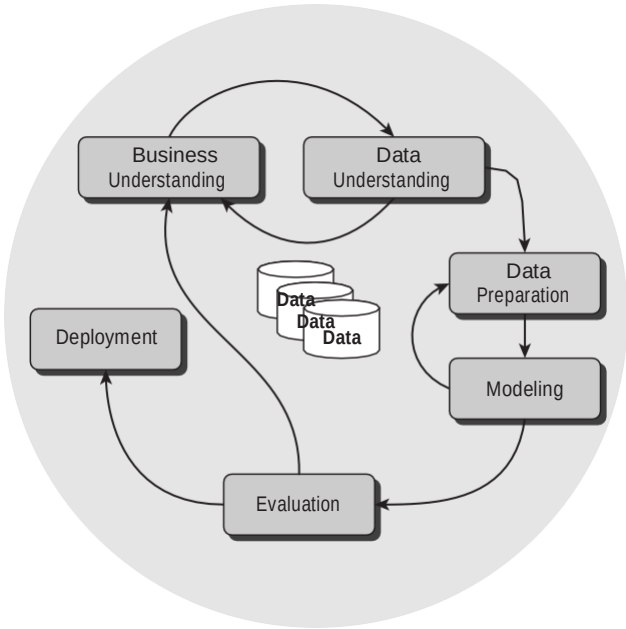

Figure 2-1: The CRISP-DM process model

Note the feedback loops in the figure. These indicate the most common ways the typical process is modified based on findings during the project. For example, if business objectives have been defined during Business Understanding, and then data is examined during Data Understanding, you may find that there is insufficient data quantity or data quality to build predictive models. In this case, Business Objectives must be re-defined with the available data in mind before proceeding to Data Preparation and Modeling. Or consider a model that has been built but has poor accuracy. Revisiting data preparation to create new derived variables is a common step to improve the models.

Business Understanding

Every predictive modeling project needs objectives. Domain experts who understand decisions, alarms, estimates, or reports that provide value to an organization must define these objectives. Analysts themselves sometimes have this expertise, although most often, managers and directors have a far better perspective on how models affect the organization. **Without domain expertise**, the definitions of what models should be built and how they should be assessed can lead to failed projects .

The Three-Legged Stool

One way to understand the collaborations that lead to predictive modeling success is to think of a three-legged stool. Each leg is critical to the stool remaining stable and fulfilling its intended purpose. In predictive modeling, the three legs of the stool are **(1) domain experts, (2) data or database experts, and (3) predictivemodeling experts**. Domain experts are needed to frame a problem properly ina way that will provide value to the organization. Data or database experts are needed to identify what data is available for predictive modeling and how that data can be accessed and normalized. Predictive modelers are needed to build the models that achieve the business objectives.

Consider what happens if one or more of these three legs are missing**. If the problem is not defined properly** and only modelers and the database adminis trator are defining the problems, excellent models may be built with fantastic accuracy only to go unused because the model doesn't address an actual need of the organization..

If the **database expert is not involved**, data problems may ensue. **First**, there may **not be enough understanding of the layout of tables** in the database to be able to access all of the fields necessary for predictive modeling. Second, there may **be insufficient understanding of fields** and what information they represent even if the names of the fields seem worse still, if the names are cryptic and no data dictionary is available. Third, **insufficient permissions may prevent pulling data into the predictive modeling environment**. Fourth, database resources **may not support the kinds of joins** the analyst may believe he or she needs to build the modeling data. And fifth, model deployment options visualized by the predictive modeling team **may not be supported bythe organization**.

If the **predictive modelers are not available** during the business understanding stage of CRISP-DM, obstacles outlined in this chapter may result. ◎ **Unrealistic Expectations**: Program managers might not fully understand what predictive models can actually do because of the hype around them. This can lead to asking for models that are impossible to build.

◎ **Poor Target Definition**: The target variables for predictive modeling might not be defined at all or might be poorly specified, which can hinder the effectiveness of the modeling efforts.

◎ **Inadequate Data Layout**: Without input from predictive modelers, the necessary data layout for building models might not be defined or may miss important fields needed for the models.

Business Objectives

Assuming all three types of individuals that make up the **three-legged stool** of predictive modeling are present during the **Business Understand stage** of CRISP-DM,.

Six key issues that should be resolved during the Business Understanding stage include definitions of the following:

- **Core business objectives** to be addressed by the predictive models
- How the business objectives can be quantified
- What data is available to quantify the business objectives
- What modeling methods can be invoked to describe or predict the business objectives
- How the goodness of model fit of the business objectives are quantified so that the model scores make business sense
- How the predictive models can be deployed operationally

Frequently, the compromises reached during discussions are the result of the imperfect environment that is typical in most organizations.

For example, data that you would want to use in the predictive models may not be available in a timely manner or at all.

Target variables that address the business objectives more directly may not exist or be able to be quantified.

Computing resources may not exist to build predictive models in the way the analysts would prefer.Or there may not be available staff to apply to the project in the timeframe needed.

And these are just a few possible issues that may be uncovered. Project managers need to be realistic about which business objectives can be achieved in the timeframe and within the budget available.

Predictive modeling covers a wide range of business objectives.. Following is a shortlist of predictive modeling projects..

PROJECT		
Customer acquisition/ Response/Lead generation	Credit card application Fraud	Medical image anomaly detection
Cross-sell/Up-sell	Loan application fraud	Radar signal, vehicle/aircraft identification
Customer next product to Purchase	Invoice fraud	Radar, friend-or-foe differentiation
Customer likelihood to purchase in *N* days	Insurance claim fraud	Sonar signal object identification (long and short range)
Website—next site to interact With	Insurance application Fraud	Optimum guidance commands for smart bombs or tank shells
Market-basket analysis	Medical billing fraud	Likelihood for flight to be on time
Customer value/Customer profitability	Payment fraud	Insurance risk of catastrophic claim
Customer segmentation	Warranty fraud	Weed tolerance to pesticides
Customer engagement with Brand	Tax collection likelihood to pay	Mean time to failure/ Likelihood to fail
Customer attrition/Retention	Non-filer predicted tax Liability	Likelihood of hardware failure due to complexity
Customer days to next Purchase	Patient likelihood to re-admit	Fault detection/Fault explanation
Customer satisfaction	Patient likelihood to comply with medication protocols	Part needed for repair
Customer sentiment/ Recommend to a friend	Cancer detection	Intrusion detection/ Likelihood of an intrusion event
Best marketing creative	Gene expression/ Identification	New hire likelihood to succeed/advance
Credit card transaction fraud	Predicted toxicity (LD50 or LC50) of substance	New hire most desirable characteristics

While many models are built to predict the behavior of people or things, not all are. Some models are built expressly for the purpose of understanding the behavior of people, things, or processes better. For example, predicting "weed tolerance to pesticides" was built to test the hypothesis that the weeds were becoming intolerant to a specific pesticide. The model identified the primary

contributors in predicting success or failure in killing the weeds; this in and of itself was insightful. While the likelihood of a customer purchasing a product within seven days is interesting on its own, understanding *why* the customer is likely to purchase can provide even more value as the business decides how best to contact the individuals. Or if a Customer Retention model is built with high accuracy, those customers that match the profile for retention

Defining Data for Predictive Modeling

Data for predictive modeling must be two-dimensional, comprised of rows and columns. Each row represents what can be called a *unit of analysis*. For customer analytics, this is typically a customer.

For fraud detection, this may bea transaction.

For call center analytics, this may refer to an individual call.

For survey analysis, this may be a single survey. The unit of analysis is problem- specific and therefore is defined as part of the Business Understanding stage of predictive modeling.

If data for modeling is loaded from files, the actual form of the data is largely irrelevant because most software packages support data in a variety of formats:

- Delimited flat files, usually delimited with commas (`.csv` files), tabs, or some other custom character to indicate where field values begin and end
- Fixed-width flat files with a fixed number of characters per field. No delimiters are needed in this format but the exact format for each field must be known before loading the data.
- Other customized flat files
- Binary files, including formats specific to software packages

Most software packages also provide connectivity to databases through native or ODBC drivers so that tables and views can be accessed directly from the software. Some software allows for the writing of simple or even complex queries to access the data from within the software itself, which is very conve-nient for several reasons:

- Data does not have to be saved to disk and loaded into the predictive modeling software, a slow process for large data sets.
- Data can be maintained in the database without having to provide version control for the flat files.
- Analysts have greater control and flexibility over the data they pull from the database or data mart.

However, you must also be careful that the tables and views being accessed remain the same throughout the modeling project and aren't changing without any warning. When data changes without the knowledge of the analyst, models can also change inexplicably.

Defining the Columns as Measures

Columns in the data are often called *attributes*, *descriptors*, *variables*, *fields*, *features*, or just columns. The book will use these labels interchangeably. Variables in the data are measures that relate to or describe the record. For customer analytics, one attribute may be the customer ID, a second the customer's age, a third the customer's street address, and so forth. The number of attributes in the data is limited only by what is measured for the particular unit of analysis, which attributes are considered to be useful, and how many attributes can be handled by the database or predictive modeling software.

Columns in the data are measures of that unit of analysis, and for predictive modeling algorithms, the number of columns and the order of the columns must be identical from record to record. Another way to describe this kind of data is that the data is rectangular. Moreover, the meaning of the columns must be consistent. If you are building models based on customer behavior, you are faced with an immediate dilemma: How can you handle customers who have visited different numbers of times and maintain the rectangular shape of the data?

Consider Table 2-2 with two customers of a hotel chain, one of whom has visited three times and the other only once. In the table layout, the column labeled "Date of Visit 1" is the date of the first visit the customer made to the hotel property. Customer 100001 has visited only once and therefore has no values for visit 2 and visit 3. These can be labeled as "NULL" or just left blank. The fact that they are not defined, however, can cause problems for some modeling algorithms, and therefore you often will not represent multiple visits as separate columns..

Table 2-2: Simple Rectangular Layout of Data

CUSTOMER ID	DATE OF VISIT 1	DATE OF VISIT 2	DATE OF VISIT 3
100001	5/2/12	NULL	NULL
100002	6/9/12	9/29/12	10/13/12

There is a second potential problem with this layout of the data, however. The "Date of Visit 1" is the first visit. What if the pattern of behavior related to the models is better represented by how the customer behaved most recently? For customer 100001, the most recent visit is contained in the column "Date of

Visit 1," whereas for customer 100002, the most recent visit is in the column "Date of Visit 3." Predictive modeling algorithms consider each column as a separate measure, and therefore, if there is a strong pattern related to the most recent visit, the pattern is broken in this representation of the data. Alternatively, you could represent the same data as shown in Table 2-3.

Table 2-3: Alternative Rectangular Layout of Data

CUSTOMER ID	DATE OF VISIT 1	DATE OF VISIT 2	DATE OF VISIT 3
100001	5/2/12	NULL	NULL
100002	10/13/12	9/29/12	6/9/12

In this data, Visit 1 is no longer the first visit but is the most recent visit, Visit 2 is two visits ago, Visit 3 is three visits ago, and so on. The representation of the data you choose is dependent on which representation is expected to provide the most predictive set of inputs to models.

A third option for this customer data is to remove the temporal data completely and represent the visits in a consistent set of attributes that summarizes the visits. Table 2-4 shows one such representation: The same two customers are described by their most recent visit, the first visit, and the number of visits.

Table 2-4: Summarized Representation of Visits

CUSTOMER ID	DATE OF FIRST VISIT	DATE OF MOST RECENT VISIT	NUMBER OF VISITS
100001	5/2/12	5/2/12	1
100002	6/9/12	10/13/12	3

Ultimately, the representation problems described in Tables 2-3, 2-4, and 2-5 occur because this data is inherently three dimensional, not two. There is a temporal dimension that has to be represented in the row-column format, usually by summarizing the temporal dimension into *features* of the data.

Defining the Unit of Analysis

Predictive modeling assumes each record is independent, meaning the algorithms don't account for connections between records, like two customers who are a married couple. If records are connected, the data won't represent general patterns accurately, leading to biased models. For example, in hospitality analytics, each customer is typically independent, but exceptions like business conventions exist.

Which Unit of Analysis?

Ultimately, the unit of analysis selected for modeling is determined by the

business objectives and how the model will be used operationally. Are decisions made from the model scores based on a transaction? Are they made based on the behavior of a single customer? Are they made based on a single visit, or based on aggregate behavior of several transactions or visits over a time period? Some organizations even build multiple models from the same data with different units of analysis precisely because the unit of analysis drives the decisions you can make from the model.