Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY.

This page intentionally left blank.

TDWI Predictive Analytics Principles and Practices

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

TABLE OF CONTENTS

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

COURSE OBJECTIVES

You will learn:

- ✓ ***Definitions, concepts, and terminology of predictive analytics***
- ✓ ***How predictive analytics relates to data science and BI programs***
- ✓ ***Structure, categories, and applications of predictive models***
- ✓ ***Enabling methods adapted from statistics, data mining, and machine learning***
- ✓ ***Proven development and implementation approaches***
- ✓ ***How human and organizational factors including team composition, structure, culture, collaboration, and accountabilities enable success***
- ✓ ***Why business, technical, and management skills are essential for success***
- ✓ ***Practical guidance for getting started with predictive analytics***

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

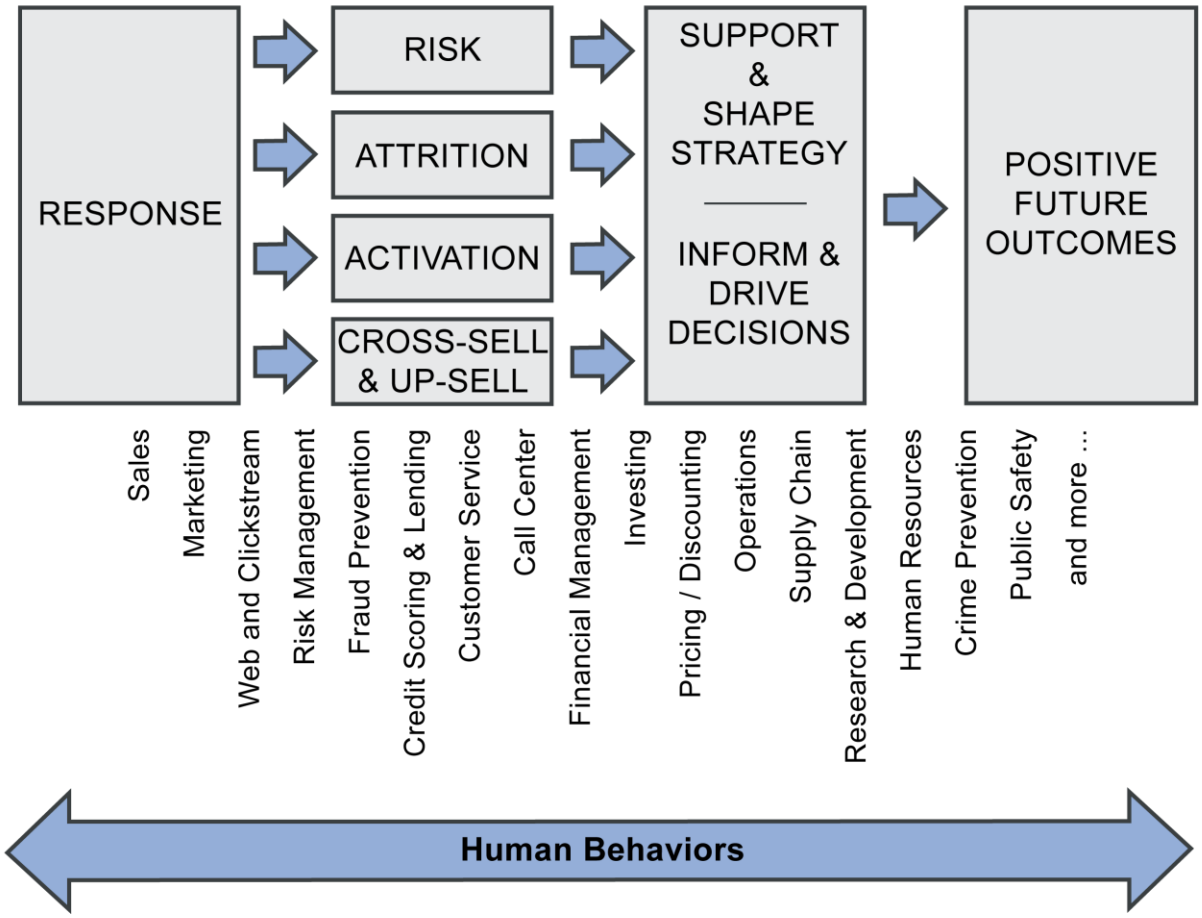
Module 1

Predictive Analytics Concepts

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Common Applications for Predictive Analytics

What Businesses Need to Predict



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Common Applications for Predictive Analytics

What Businesses Need to Predict

RESPONSE PREDICTION

Predictive analytics is predicated on the concept of predicting human behaviors—what people will do in specific circumstances. Every predictive analytics project begins with response prediction where the goal is to understand how people (and various segments of a population) will respond to a specific situation or stimulus.

EXTENDING THE RESPONSE MODEL

Response predictions are typically extended or adapted to specific needs and circumstances. A response model may be extended to predict:

- Risk—predictions about segments of a population that may engage in fraud, commit crimes, compromise workplace safety, etc.
- Attrition—predictions about segments of a population that may be lost as customers, employees, contributors, partners, etc.
- Activation—predictions of probability (by segment) to set a process in motion, such as activating a trial version of a software product
- Cross-sell and up-sell—predictions of probability that purchasers will respond to suggestions for related products and services

APPLYING THE RESPONSE MODEL

As already discussed, the value of predictive analytics is achieved by applying the predictions to support and shape strategy and to inform and drive decisions.

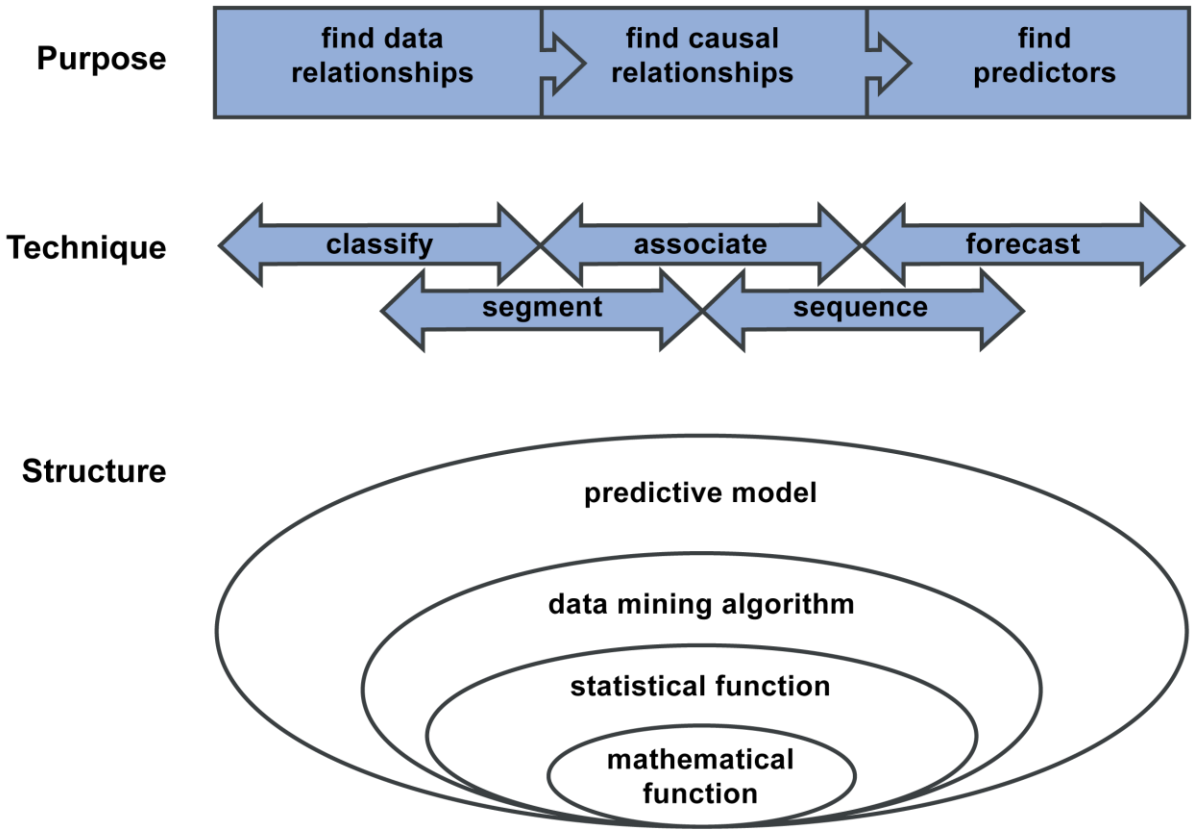
BUSINESS OUTCOMES

Predictions of response, extended to context of business need and used to drive positive business outcomes, are the keys to effective predictive analytics. Positive business outcomes are specifically related to business domains. The facing page illustrates many of the common business domains—from sales and marketing to public safety—where value is created with predictive analytics.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

The Language of Predictive Analytics

Making Sense of the Terminology



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

The Language of Predictive Analytics

Making Sense of the Terminology

COMMON WORDS WITH SPECIFIC MEANINGS

The language of data mining and predictive analytics can be confusing to those who are just becoming familiar with the field. In many cases commonly used words have very specific meanings in data mining and predictive modeling contexts.

PURPOSE

A data mining model typically serves to find one of three things:

- Data relationships—hidden patterns in the data that are useful for gaining new knowledge and understanding
- Causal relationships—relationships that indicate cause and effect
- Predictors—the variables in a relationship that are useful to predict future outcomes.

TECHNIQUE AND STRUCTURE

Data mining uses multiple techniques to achieve different objectives. These techniques, discussed in greater depth later in this course:

- Classify—find groups of similar objects in a set of data
- Segment—divide data into subsets based on the groups found through classification
- Associate—find relationships among variables (attributes) in a set of data
- Sequence—discover patterns in the data where sequence or order of occurrence is apparent across multiple events
- Forecast—use historical data to understand future trends and probabilities

STRUCTURE

A data mining model is a combination of data and software that is used to draw inferences from data relationships and to generate predictions about the subjects of the data. The software applies algorithms to data to implement selected data mining techniques.

- A *model* is built using one or more algorithms.
- An *algorithm* applies one or more functions to determine a result.
- A *statistical function* applies one or more mathematical functions to derive statistical values. Some statistical functions called *distributions* generate a range of possible values for a given input value. These are discussed in more detail later in the course.
- A *mathematical function* is an equation or formula that generates an output value based on or more input values.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Module 2

Models and Statistics

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Models

What Are Models?

Models

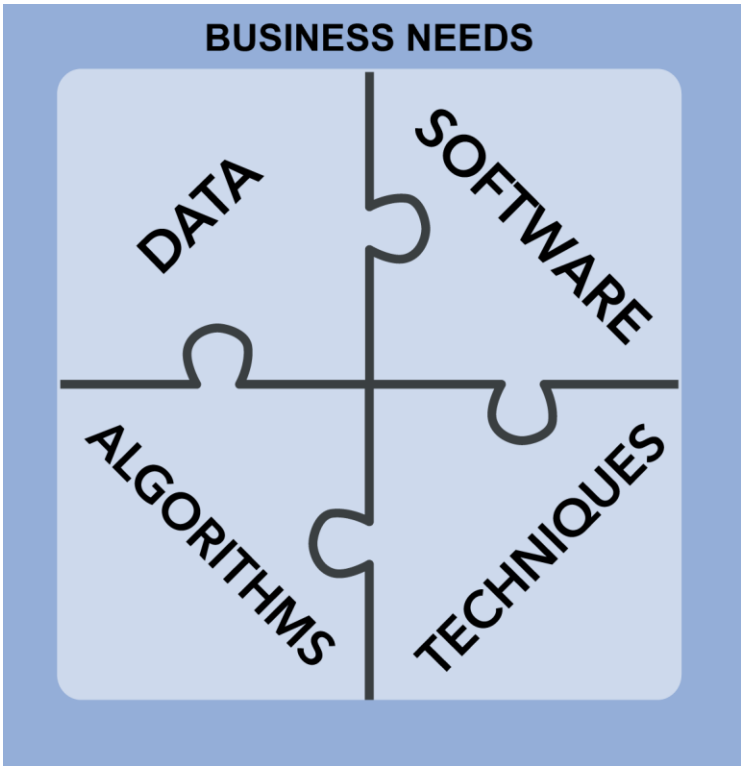
- Simplified representations of reality that meet defined information needs

Main Components of Models

- Data
- Software
- Algorithms
- Techniques

Other Components of Models

- Symbols
- Rules
- Diagrams
- Physical items



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Models

What Are Models?

MODELS

Models are simplified abstractions of reality. By design, models are simplified to a certain level of detail that still maintains their usefulness and relevance to a given problem while eliminating detail that is not relevant.

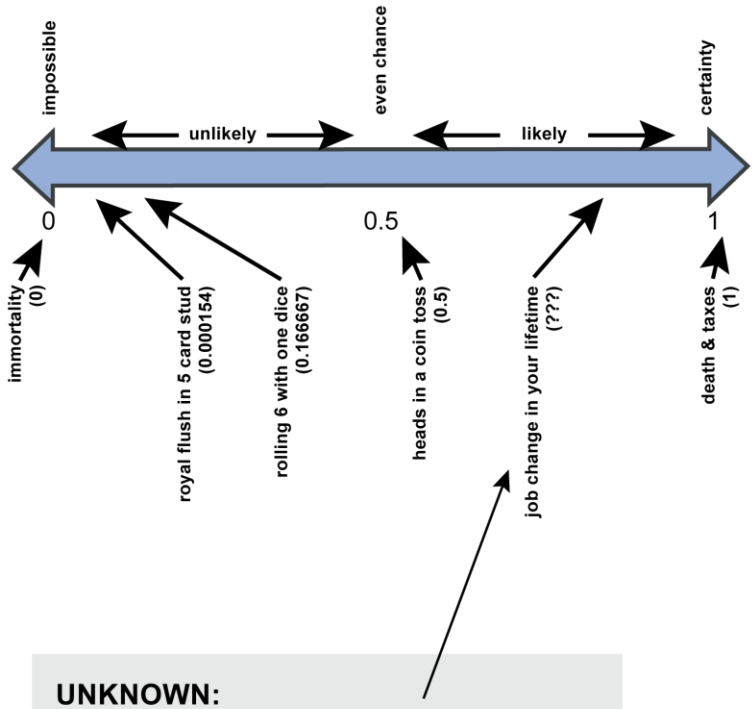
An *analytical model* combines data with software to help draw inferences from the data and to create predictions for some of the attributes contained in the data. The software applies algorithms to the data to implement selected data mining, statistical, or machine learning techniques.

The selection of data, algorithm, and technique for a given situation is driven by the problem context and by the needs people have for the model's informational output. Models are the building blocks used by automated processes that search for patterns and relationships in large data sets.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Probability

Estimating Likelihood



UNKNOWN:
Job change probability is unknown.
How can we analyze it?

Business Context
What are the goals of analysis?

Data
What data is needed?
What are the characteristics of the data?
What does the data tell us?
(e.g., is job change an independent or dependent event?)

Analytic Modeling
Which techniques and how to apply them?

Evaluation
Are the analysis goals met?
Have we measured probability of job change?
Is it a useful and reliable predictor?

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Probability

Estimating Likelihood

MEASURING PROBABILITY

Probability is an important concept in predictive analytics. FICO, the decision management company best known for credit scoring, says that "predictive analytics turns uncertainty into usable probability." In statistical terms, probability measures how likely it is that something will occur. The value of the measure is always in the range of zero to one, where zero corresponds with impossible and one corresponds with complete certainty. The scale on the facing page illustrates several examples of probability between the two extremes.

ANALYZING PROBABILITY

Probability analysis is investigation and study to turn uncertainty into a usable probability measure. Many of the examples shown here are mathematically certain—probability of heads on a coin toss is always 0.5. One of the examples—job change—is uncertain and a good candidate for analysis. Statistical analysis and data mining provide the means to perform that analysis, beginning with problem context and ending with a useful probability measure.

PROBABILITY MODELS

There are many techniques that help us create models that calculate probabilities. Some of the terms you may encounter include:

- Statistical models
- Distribution models
- Data mining techniques
- Machine learning algorithms

In statistics, there are multiple ways to estimate probabilities. This module explores probability distribution models. Module 3 will explore regression models. In predictive analytics, additional techniques include data mining and machine learning. These will be discussed in Module 4.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Transforming Data With Intelligence™

Module 3

Regression Model Examples

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Linear Regression Models

Overview

Linear Regression Models
Statistical models that estimate relationships between a continuous dependent variable and one or more independent variables.

- Wikipedia

Areas of Applicability
- Use historical data to create models that forecast and predict values of continuous variables based on input independent variables.
- Single-input models are called simple linear regression models.
- Multi-input models are called multiple linear regression models.

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Linear Regression Models

Overview

LINEAR REGRESSION

Linear regression models calculate continuous dependent variables to predict or forecast their future values.

A linear regression model is used when the dependent variable is continuous and the independent variables are continuous. The technique can also support ordinal and nominal (categorical) independent variables if they can be transformed to numerical values.

Linear regression models are created using statistical techniques. These techniques vary in how they arrive at a formula (or model) that best describes the observed data. Regression techniques attempt to minimize the error between the observed data and the generated model. The statistics and machine learning communities provide collaboration and contribution to development of algorithms in this area.

APPLICATIONS

Regression techniques use historical data to estimate the parameters of a model that may be used to forecast and predict future values of the dependent variable.

Simple linear regression models are developed to study the influence of a single independent variable on a single dependent variable. An example is provided on the following pages.

Multiple linear regression models may be developed to study the collective influence of several independent variables on a continuous dependent variable. When multiple independent variables are considered simultaneously, the interaction effects of the independent variables influencing the dependent variable can be studied and analyzed. This type of model helps managers make informed trade-off decisions by adjusting the decision variables in a manner that produces an acceptable output value.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Logistic Regression Models

Overview

Logistic Regression Models
Statistical models that estimate relationships between a binary dependent variable and one or more independent variables. Used for predicting and classifying outcomes.

- Wikipedia

Areas of Application

- Use historical data to create models that calculate probability values for a binary condition or event to occur
- Independent variables can be continuous, nominal (categorical) or ordinal
- Single-input models are called simple logistic regression models
- Multi-input models are called multiple logistic regression models

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Logistic Regression Models

Overview

LOGISTIC REGRESSION

Logistic regression is used when the dependent variable is a binary category. Examples of binary categories include true vs. false, yes vs. no, win vs. lose, and so forth. Logistic regression produces a probability for the desired outcome based on the values of independent variables.

The independent variables can be continuous, ordinal, or nominal. When ordinal or nominal independent variables are used, they must be transformed to numerical values.

Logistic regression models use the logit function in producing output variables. Typically, the model produces the log of the odds ratio, then transforms it into a useful probability value.

The following pages provide a simple example of logistic regression with one independent variable. As with linear regression, there are multiple techniques for logistic regression. The statistics and machine learning communities provide collaboration and contribute to development of algorithms in this area.

APPLICATIONS

Logistic regression may be used to calculate the probability of future events based on historical data. This is a fundamental capability for predictive analytics.

A logistic regression model is essentially a calculator that is able to estimate probabilities resulting from combinations and interactions of multiple independent variables. These models are complementary to distribution models discussed previously.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Transforming Data With Intelligence™

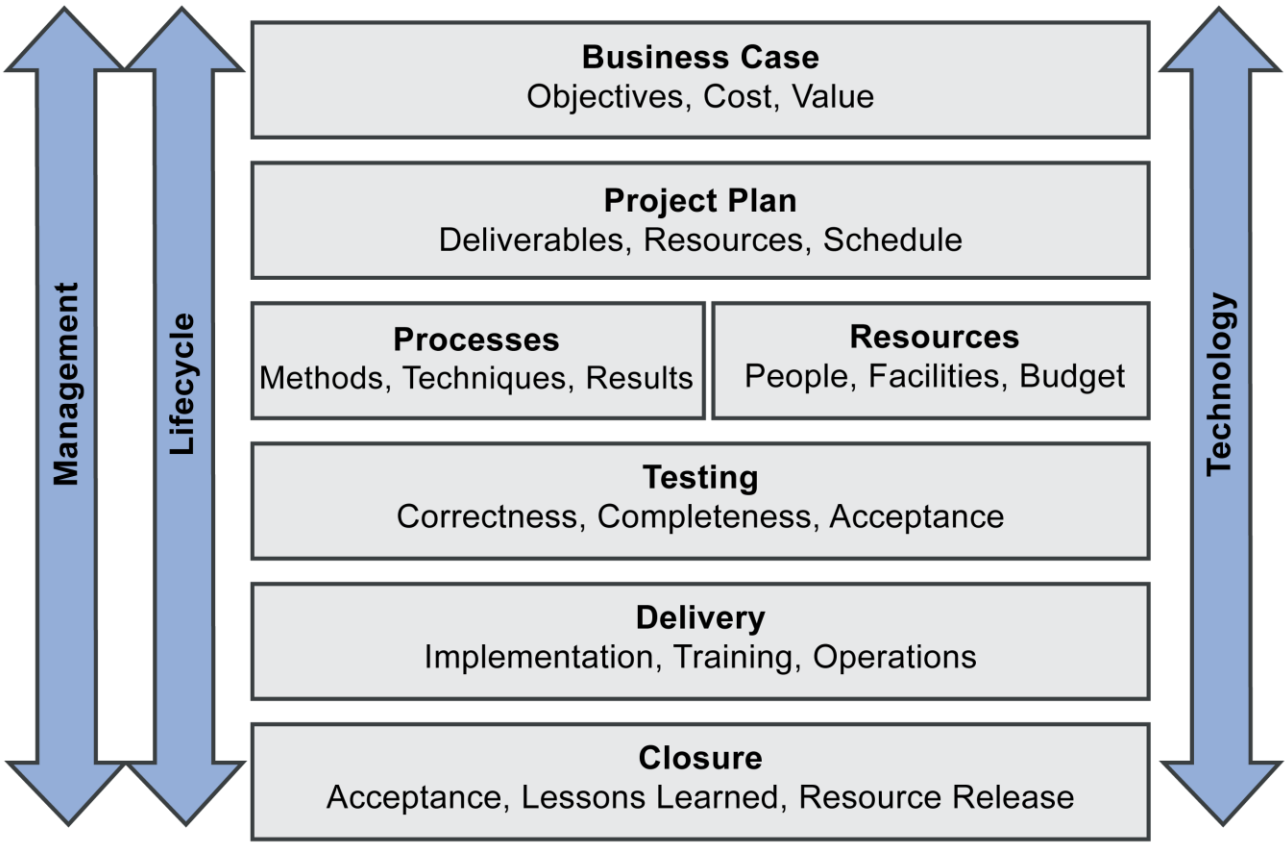
Module 4

Building Predictive Models

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Model Building Processes

Data Mining Projects



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Model Building Processes

Data Mining Projects

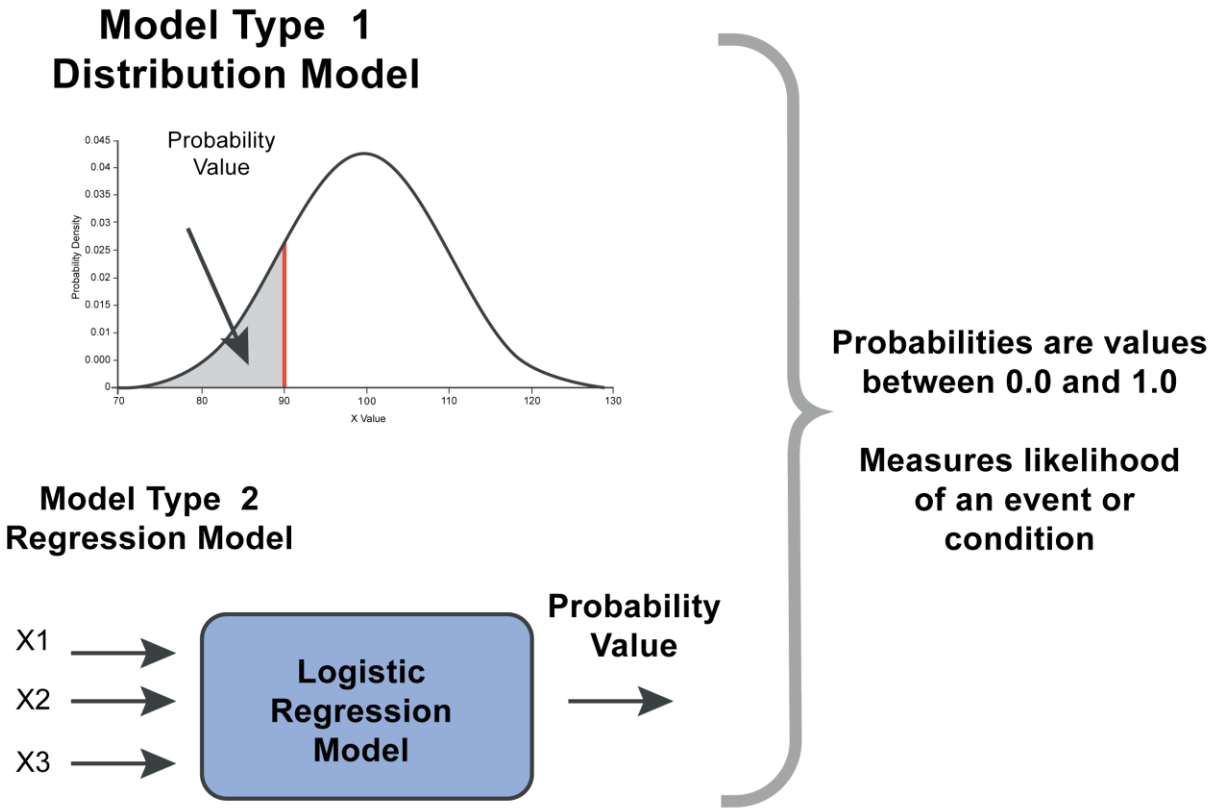
PROJECT DISCIPLINE

Data mining is complex and should be undertaken with the discipline of projects. Avoid the temptation to "jump into the deep end of the data pool," and start with business objectives and benefits. Plan the project before committing resources and performing data mining processes and activities. Test results and then deliver. Finally close the project with formal acceptance of deliverables and a project review. Execute data mining projects with the right level of project management, a defined lifecycle, and the right technology to do the job well.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Techniques

Probability Values



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Techniques

Probability Values

FOUNDATION OF PREDICTION

As described in an earlier section of the course, a probability value is a measure of the likelihood of a future event of condition. It provides the most fundamental form of a prediction.

Probabilities are commonly made at the granularity of an event. Other forms of predictions described in the module consider other forms of granularity of the prediction.

Common modeling approaches for creating probability estimates include distribution models and logistic regression models.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Transforming Data With Intelligence™

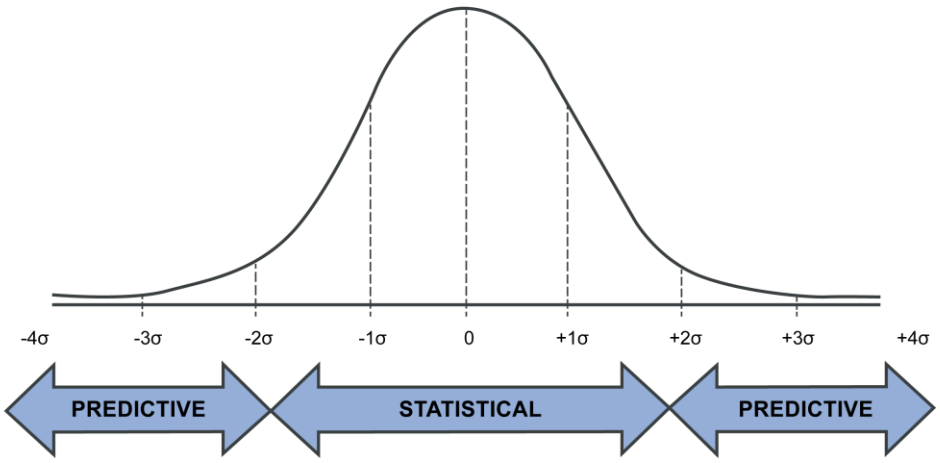
Module 5

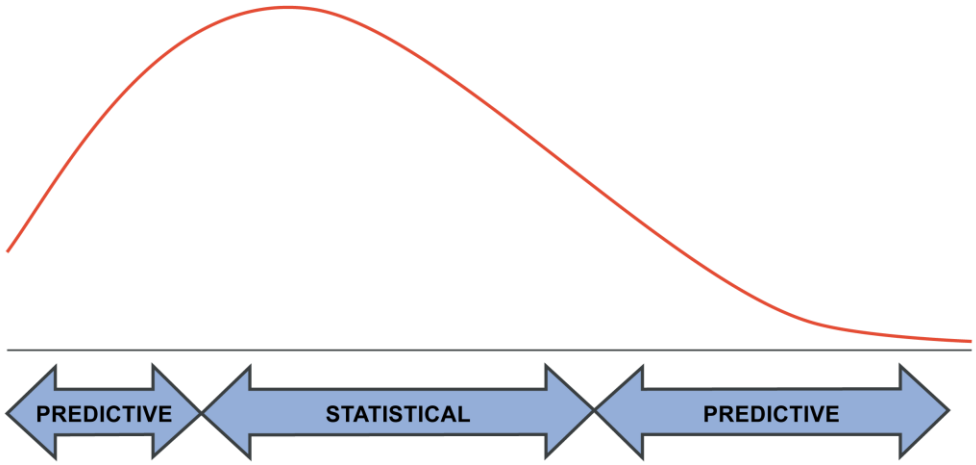
Implementing Predictive Capabilities

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Introductory Concepts

Distribution View





 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Introductory Concepts

Distribution View

STATISTICS AND DISTRIBUTION

Traditional statistical analysis is primarily focused on central tendencies—the center of the distribution curve. As variation and standard deviation increase, the information and analytics value declines. This works because the analysis is centered on understanding the nature of outcomes.

PREDICTIVE WITH NORMAL DISTRIBUTION

Predictive analytics shifts the attention away from central tendencies to look at the tails of the curve and things that are distant from central tendencies. In predictive analytics, the purpose is not to understand the nature of outcomes, but to shape future outcomes. Opportunities to enhance business performance are found in the low-incidence, high-impact occurrences in the tails of the distribution. To enhance business performance we must look outside the norm.

PREDICTIVE WITH SKEWED DISTRIBUTION

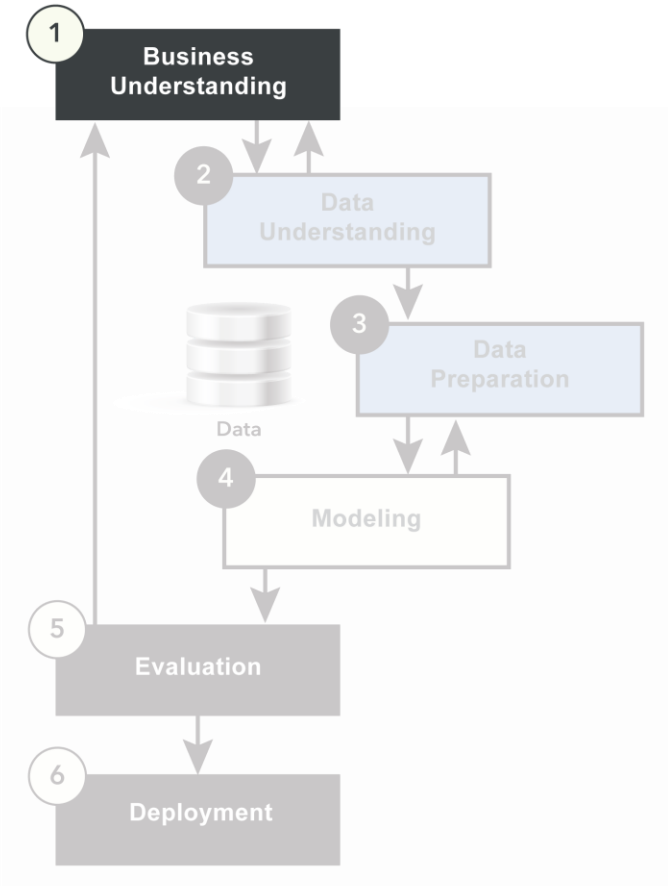
Although normal distribution is an important and central concept to analytics, business is not distributed normally in the real world. In predictive analytics (in fact, in all analytics) we must work with skewed distributions.

With the skewed distribution, both tails are still the focus. The longer tail, however, may be the most rewarding. As a practical matter, it is more difficult for predictive modeling to succeed in the tail closest to the mode due to proximity. When successful, it often yields lower impact than a long tail. The reduction in individual behavior impact, however, is partially offset by higher frequency of occurrences in this tail.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Business Understanding

Activities and Deliverables



1.1 Determine Business Objectives
- **1.1.1 Background**
- **1.1.2 Business Objectives**
- **1.1.3 Business Success Criteria**

1.2 Assess Situation
- **1.2.1 Resources Inventory**
- **1.2.2 Requirements, Assumptions, Constraints**
- **1.2.3 Terminology**
- **1.2.4 Risks & Contingencies**
- **1.2.5 Costs & Benefits**

1.3 Determine Data Mining Goals
- **1.3.1 Data Mining Goals**
- **1.3.2 Data Mining Success Criteria**

1.4 Produce Project Plan
- **1.4.1 Project Plan**
- **1.4.2 Initial Tools & Techniques Assessment**

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Business Understanding

Activities and Deliverables

BUSINESS OBJECTIVES

Start the process with business and customer perspective. Understanding what the sponsor or customer wants to accomplish is essential to avoid the risk of building elegant models with little or no value. Deliverables from this task include:

Background	Any information about the business circumstances that may affect or be useful to inform the project.
Business Objectives	A description of what the customer wants to accomplish including decisions to be made and questions needing answers.
Business Success Criteria	The measures that will be used to judge usefulness of project results. For subjective criteria, know how and by whom the criteria will be judged.

ASSESS SITUATION

Identify resources, constraints, assumptions, and other factors that influence the analysis goals and that must be considered to develop an objective and realistic project plan. Deliverables include:

Resources Inventory	A list of people, data, tools, and technology available for the project.
Requirements, Assumptions, and Constraints	Project requirements including schedule, quality, and security; assumptions about business and data; resource and technology constraints.
Terminology	A glossary of important language for the project including both business terms and data mining terms.
Risks & Contingencies	A list of factors that might cause project delay or failure, along with mitigation and contingency plans.
Costs & Benefits	An assessment of cost-effectiveness of the project weighing anticipated value against estimated costs.

DETERMINE DATA MINING GOALS

Extend from business objectives (in business terms) to describe the project goals in technical terms. Deliverables include:

Data Mining Goals	A description of project outputs and their relationships to the business objectives.
Data Mining Success Criteria	Technical success criteria for the project such as information quality and predictive accuracy.

PRODUCE PROJECT PLAN

Develop a project plan that accounts for iteration and includes initial view of mining tools and techniques. Deliverables include:

Project Plan	Typical phase-by-phase plan and schedule for the project.
Tools & Techniques Assessment	Initial concept of data mining techniques to be used and the choice of tools to apply those techniques.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Transforming Data With Intelligence™

Module 6

Human Factors in Predictive Analytics

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Culture

Executive Buy-In

 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Culture

Executive Buy-In

CULTURE DEFINED

In business, culture relates to the shared values, attitudes, standards, and beliefs that characterize the people in an organization and define the nature of that organization. Organizational culture is rooted in goals, strategies, structure, and relationships with employees, customers, stakeholders, and community. Culture is an essential component in success or failure of many business endeavors including business analytics.

SHAPING ANALYTICS CULTURE

"Culture always starts with the owner."[1] This quote captures the essence of analytics sponsorship and the importance of executive buy-in. When the executives demonstrate trust in analytics, that trust tends to permeate through the layers of the organization. When they visibly and vocally support analytics as a core decision-making and management competency, then others follow with open support. When resources for analytics are provided at the top, then middle and line managers tend to find budget, people, and time to dedicate to their local analytics needs.

[1] Patrick, Josh [2013]. "The Real Meaning of Corporate Culture," *New York Times*, May 21. http://boss.blogs.nytimes.com/2013/05/21/the-real-meaning-of-corporate-culture/?_r=0)

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Transforming Data With Intelligence™

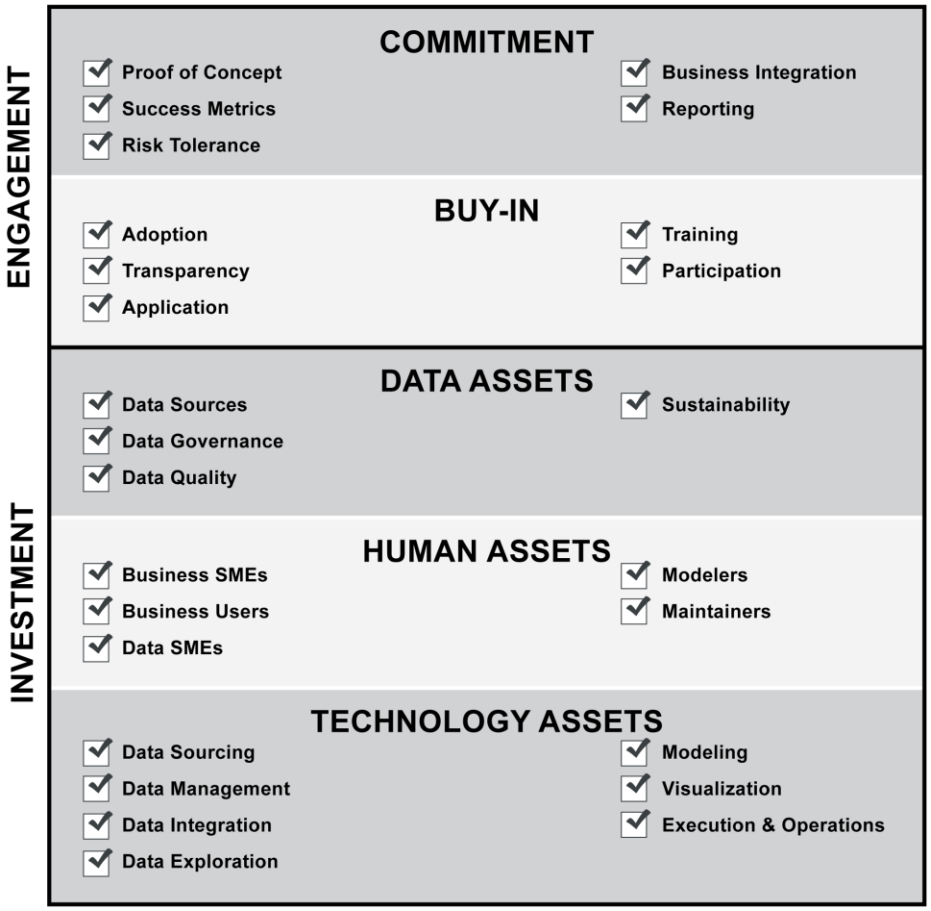
Module 7

Getting Started with Predictive Analytics

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Analytics Readiness

Readiness Checklist



 © TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Predictive Analytics Readiness

Readiness Checklist

ASSESSING THE CURRENT STATE

Think of predictive analytics as a journey along the path of maturing organizational intelligence and decision capabilities. As with any journey it is important to understand your current position before taking the first step. The readiness checklist on the facing page describes several categories to consider when evaluating your current position. Getting started in the right way is primarily about determination and the assets that you have to enable success—engagement in the form of commitment and buy-in that is supported with data, human, and technology assets.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY