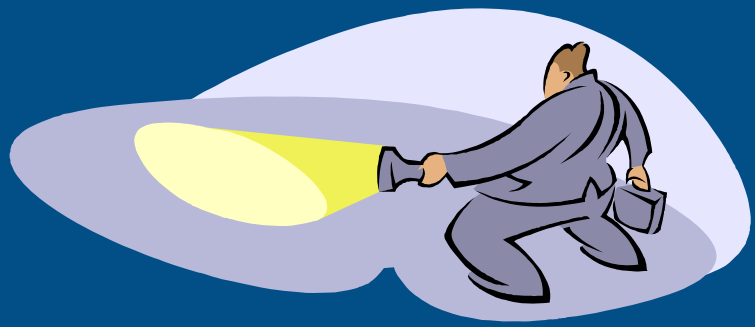
Using Decision Tree Analysis In Enrolment Management



Don Curzon BA, MBA
Director of Institutional Research
Sheridan Institute of Technology and Advanced Learning

INTRODUCTION

Don Curzon

Director of Institutional Research

Sheridan Institute of Technology and Advanced Learning

donald.curzon@sheridaninstitute.ca 905-845-9430 x 2396

- MBA - Wilfrid Laurier
- BA - York and Guelph (Personality and Behaviour)
- 10 years with Globe and Mail – Market Analytics
- 2 years with Manulife Financial – MIS
- 4 years with GlaxoSmithKline – Analytics
- 2 years with Anderson DDB – CRM Research

Agenda

1. Enrolment Management – the Need
2. The Data
3. Decision Tree Analysis – the Opportunity
4. The Model
5. The Findings
6. The Benefits

1. Enrolment Management – the Need

- **Mission-critical – funding, fixed costs**
- **Long, detailed, multi-deadline, time-sensitive yearly cycle**
- **Volatile goals - targets can change, student application activity can surprise**
- **Many constituents – government, student, industry, parent, unions, high school counsellor, Deans, Finance, Facilities**
- **Competition and markets**
- **Capital-intensive, immovable objects (space, faculty supply)**

1. Enrolment Management - Complexity

I have many offers out.

I have many seats to fill.

The clock is ticking…

Am I going to make my overall target?

Which programs are OK, and which are in danger?

Will those offered their second or third choice go elsewhere?

Does going elsewhere depend on where the applicant lives?

Does going elsewhere depend on the program?

Will gender or age affect acceptance?

Is everything going to work out in the next month, or is it already too late to act?

MORE COMPLEXITY

If I could answer each of those questions…

- How do I wade through all the permutations without missing anything?
- How do I evaluate which one is most important?
- If two factors are acting together, how do I figure out how much more important that is?
- Will I have to do this all again in April?

Is there something important I haven't considered?

2. The Data

- **Systems pretty stable**
  - **4 years of consistent data collection in a stable new system**
- **Sheridan is a large player**
  - **Student counts high enough to produce statistical significance**
- **The data is rich and complex**
  - **Rich application and applicant details and all transactions over 10 months**

3. The Opportunity

- **Lots of data, with width and length, to work with**
- **The rules are clear and measurable**
- **Acceptance is a binary variable, well-suited to decision tree analysis**
- **The work mostly comprises high task density and knowable but shifting dynamics**

PREDICTIVE MODELING

Predictive modeling tries to find good rules (models) for guessing (predicting) the values of one or more variables in a data set from the values of other variables in the data set.

Once a good rule has been found, it can be applied to new data sets that may or may not contain the variable(s) being predicted.

Neural Nets, Decision Trees, Regression

Sheridan

Neural Nets

involve network of simple processing elements (neurons), which can exhibit complex global behaviour, determined by the connections between the processing elements and element parameters.

Decision Trees

used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.

Regression

a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the *regression equation* is modeled as a function of the independent variables, corresponding parameters ("constants"), and an *error term*.

-Wikipedia

Cluster Analysis

Clustering - the process of dividing a data set into mutually exclusive groups such that the observations for each group are as close as possible to one another, and different groups are as far as possible from one another.

-SAS

Predictive Modeling vs. Clustering

Sheridan

Predictive modeling identifies a variable that needs to be explained, then uses modeling methods to predict the variable using other attributes

Cluster analysis does not identify a target variable. It uses attributes in each observation to identify observations that are most like each other and least like other observations. It does not predict.

PREDICTIVE MODELING

Sheridan

What problem does predictive modeling solve?

Why should I care?

We shall return…

4. The Model

Enrolment Management

- **Predictive modeling processes using Decision Tree Analysis to predict probability of acceptance of an offer**

Please note: This example assumes the completion of a ranking of applicants based on admission requirements, and offers being made in accordance with Ontario regulations and ethical practices. This example is about predicting acceptance and benefiting from it.

A DECISION TREE

Sheridan

The Purpose:

- Predict who will accept an offer

The Benefit:

- Know early if we've made sufficient effort
- If we need more, know how much more effort we need to make

A DECISION TREE

Setting up a decision tree model…

Identify any potentially explanatory data

Choose target variable (Fees paid)

Assess and include/exclude variables

Data Partition

Data | Variables | Partition | Stratification | User Defined | Output | Notes

Name	Model Role	Measurement	Type	Format	Label
UNIQUE	rejected	nominal	char	$23.	Unique
FEES_PAID	target	binary	char	$3.	Fees Paid
STD_ID	rejected	nominal	char	$9.	STD ID
ADMIT_TERM	rejected	unary	char	$4.	Admit Term
TERM_DESCR	rejected	unary	char	$9.	Term Descr
APPL_DT	rejected	nominal	char	$10.	Appl Dt
APPL_NBR	rejected	nominal	char	$8.	Appl Nbr
PROG_NBR	input	nominal	char	$1.	Prog Nbr
SEQUENCE	rejected	nominal	char	$1.	Sequence
ACAD_PLAN	rejected	nominal	char	$5.	Acad Plan
PLAN_DESCR	input	nominal	char	$30.	Plan Descr
CAMPUS	input	nominal	char	$3.	Campus
ACAD_LOAD	input	binary	char	$1.	Acad Load
RESIDENCY	input	nominal	char	$4.	Residency
RESD_DESCR	rejected	nominal	char	$15.	Resd Descr
PROG_STATUS	input	nominal	char	$2.	Prog Status
PROG_ACTN	rejected	binary	char	$4.	Prog Actn
PROG_ACTN_DESCR	input	binary	char	$11.	Prog Actn Descr
ACTION_RSN	rejected	nominal	char	$2.	Action Rsn
ACTN_RSN_DESCR	input	nominal	char	$23.	Actn Rsn Descr
ACTION_DT	rejected	nominal	char	$10.	Action Dt
DAYS_AFTER_JAN_1	input	interval	num	BEST12.	Days after Jan 1
MONTH_OF_OFFER	input	interval	num	BEST12.	Month of Offer

A DECISION TREE

Sheridan

What happens when you run a decision tree model?

Software determines each variable's ability to explain the target

Data Partition

Data | Variables | Partition | Stratification | User Defined | Output | Notes

Name	Model Role	Measurement	Type	Format	Label
UNIQUE	rejected	nominal	char	$23.	Unique
FEES_PAID	target	binary	char	$3.	Fees Paid
STD_ID	rejected	nominal	char	$9.	STD ID
ADMIT_TERM	rejected	unary	char	$4.	Admit Term
TERM_DESCR	rejected	unary	char	$9.	Term Descr
APPL_DT	rejected	nominal	char	$10.	Appl Dt
APPL_NBR	rejected	nominal	char	$8.	Appl Nbr
PROG_NBR	input	nominal	char	$1.	Prog Nbr
SEQUENCE	rejected	nominal	char	$1.	Sequence
ACAD_PLAN	rejected	nominal	char	$5.	Acad Plan
PLAN_DESCR	input	nominal	char	$30.	Plan Descr
CAMPUS	input	nominal	char	$3.	Campus
ACAD_LOAD	input	binary	char	$1.	Acad Load
RESIDENCY	input	nominal	char	$4.	Residency
RESD_DESCR	rejected	nominal	char	$15.	Resd Descr
PROG_STATUS	input	nominal	char	$2.	Prog Status
PROG_ACTN	rejected	binary	char	$4.	Prog Actn
PROG_ACTN_DESCR	input	binary	char	$11.	Prog Actn Descr
ACTION_RSN	rejected	nominal	char	$2.	Action Rsn
ACTN_RSN_DESCR	input	nominal	char	$23.	Actn Rsn Descr
ACTION_DT	rejected	nominal	char	$10.	Action Dt
DAYS_AFTER_JAN_1	input	interval	num	BEST12.	Days after Jan 1
MONTH_OF_OFFER	input	interval	num	BEST12.	Month of Offer

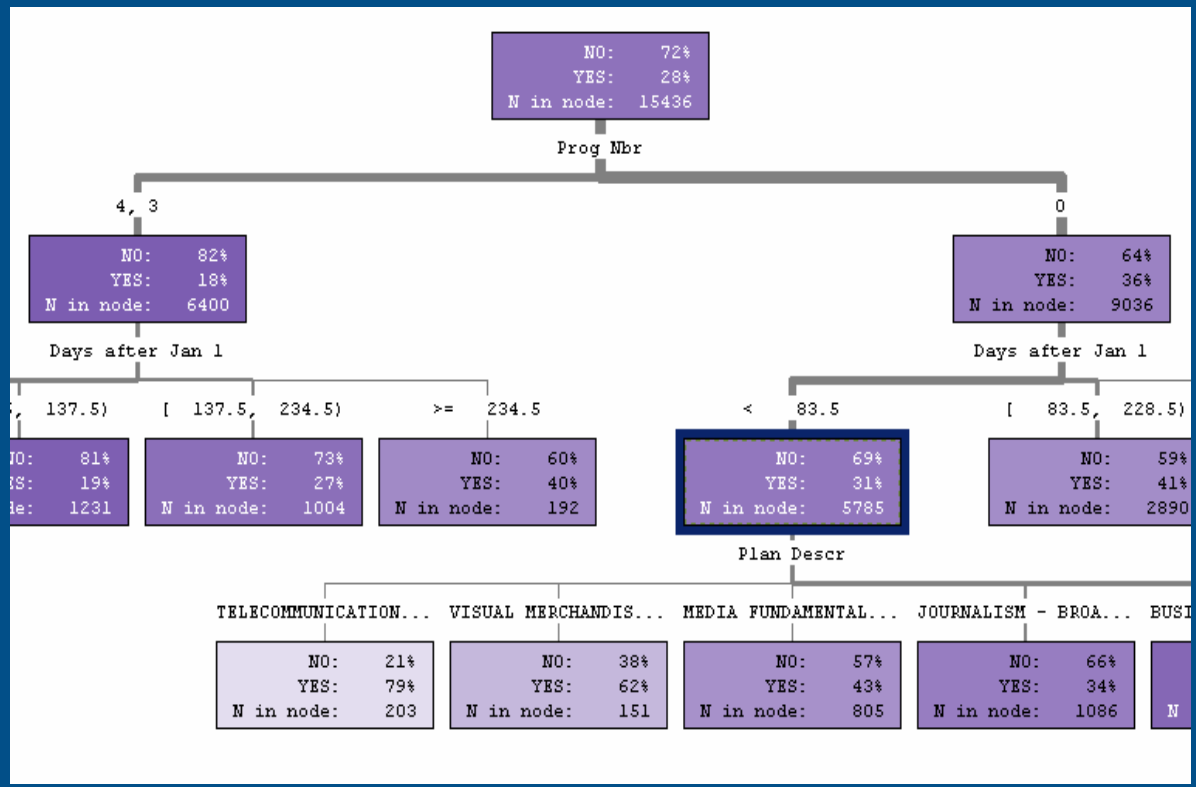
A DECISION TREE

The visual representation:

The tree has nodes (boxes), branches, levels, and leaves (endpoints).

The top node is the full dataset. Each branch divides the data based on predictors of offer acceptance.

Those divisions are found by predictive analysis of the explanatory variables.



A DECISION TREE

The top node:

- 15,436 offers
- 72% did not accept
- 28% did accept

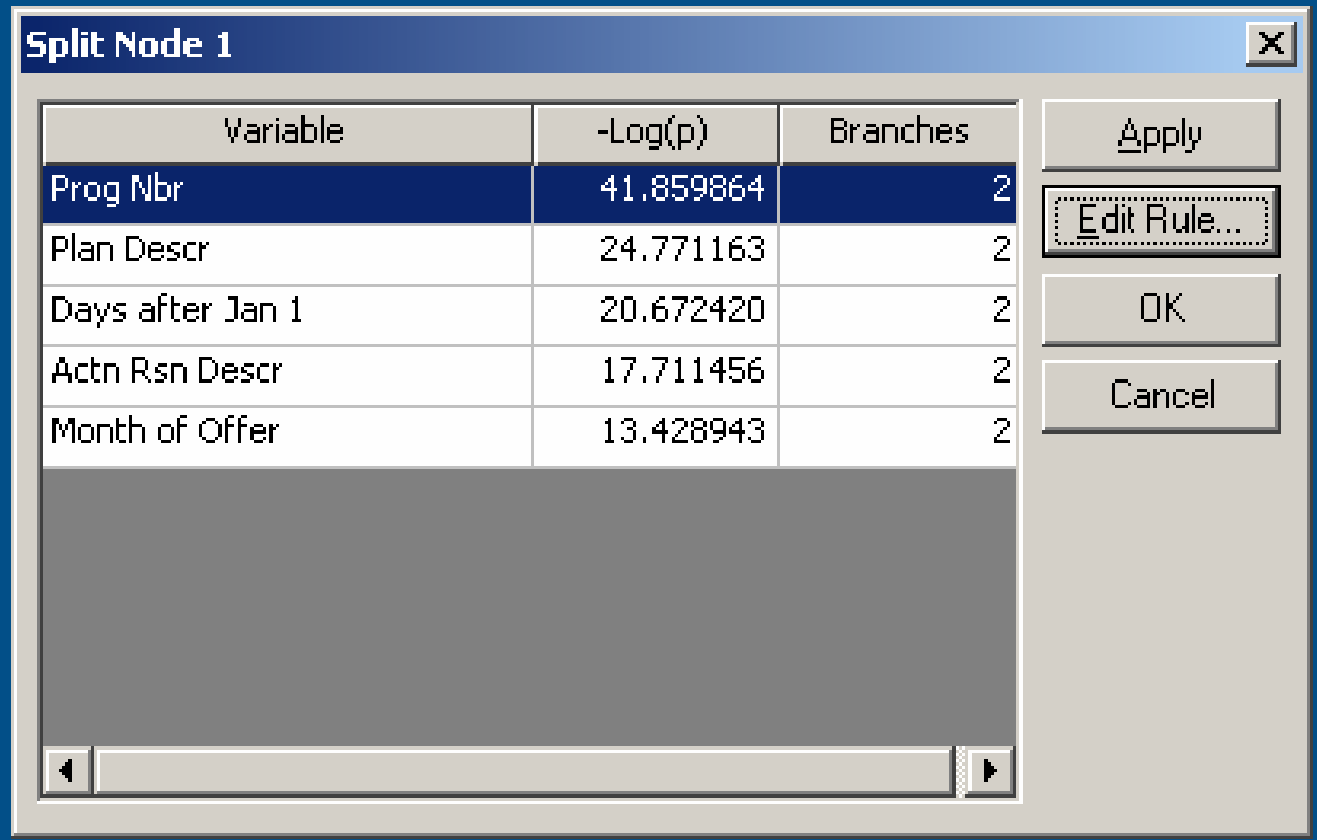
NO:	72%
YES:	28%
N in node:	15436

A DECISION TREE

Compared and Ranked:

Variables are ranked by predictive power

The winner drives the first split in the tree, in this case the applicants' preferences or Choices (Prog Nbr)



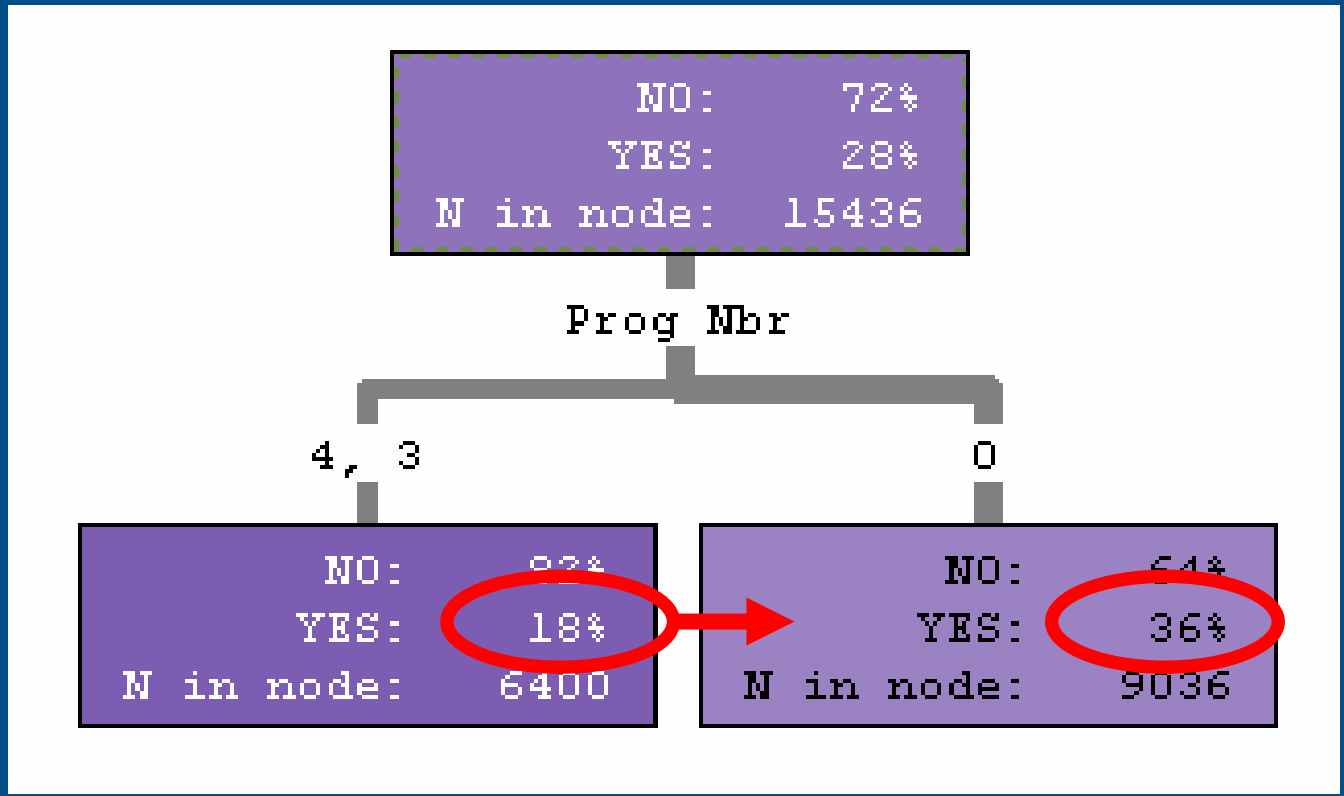
A DECISION TREE

The first split:

Choice (Prog Nbr) is the best predictor

When offers are for 1st choice, acceptance doubles

Data is now divided into 2 nodes based on the importance of Choice



A DECISION TREE

The second comparison and ranking:

For the first split (Choice) the next-best predictor is how late in the year it is when offers were accepted

Note that the Program Applied To (Plan Descr) is a close second

Split Node 102

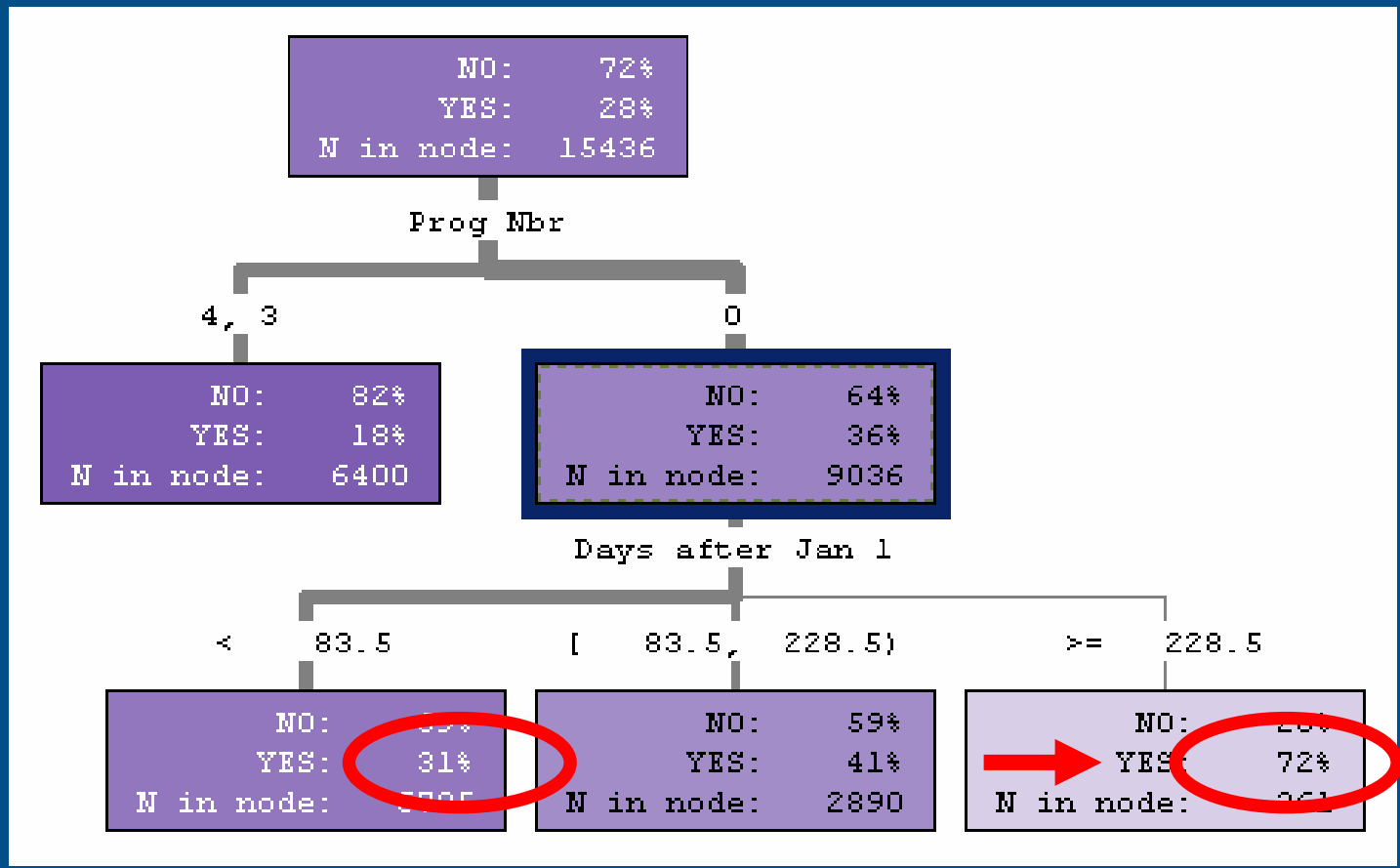
Variable	-Log(p)	Branches
Days after Jan 1	38.005549	3
Plan Descr	33.520700	5
Month of Offer	30.106530	5
Actn Rsn Descr	24.453656	4
Campus	10.394029	2

Apply
Edit Rule...
OK
Cancel

A DECISION TREE

The second split, for 1st Choice:

The later it is, the more likely the applicant will accept (the acceptance rate more than doubles)



A DECISION TREE

Examples of discoveries:

- Getting first choice doubles likelihood of accepting
- Last minute applicants who get their first choice double again in likelihood of accepting
- Highly respected programs attract high acceptance rates, especially if first choice

A DECISION TREE

Sheridan

Let’s open up a model

and see how it works…

PREDICTIVE MODELING

Sheridan

What problem does predictive modeling solve?

Why should I care?

Some answers…

WE NOW KNOW…

• The most important factors	• Choice, deadlines, program
• Order of importance	• Top-to-bottom branching
• How much each one matters	• Probability, error reduction, target change
• How they interact	• Choice then deadlines then…
• What's unimportant	• Had little or no predictive value
• Easiest to act on	• High probability applied to biggest gap

A COMPARISON OF PROCESSES

Traditional Method

- Observation
- Consultation/Intuition
- Hypothesis
- Collect, crosstab, corre
- Anything interesting?

Outcome 1

- Found something
- Spent money
- Not sure what happened
- Repeat

A COMPARISON OF OUTCOMES

Sheridan

Predictive Modeling

- Choose target
- Kitchen sink
- Push button
- See everything

Outcome 2

- Know importance
- Know interactions
- Act, see effects
- Refresh the data & repeat

BENEFITS OF MODELING

Sheridan

- "Kitchen sink" can go into the model
- No assumptions required
- No intuition about where to spend precious time – it's not a sequential process
- Tactics and strategies are obvious, in fact can be priced within the model
- Find things you didn't think of
- Build a scenario…and another…

CAUTIONS

- “The Kitchen Sink” - overfit and colinearity
- The right method for the job
- Causations, dependencies, declining returns
- What’s missing?
- Take lessons

CAUTIONS ILLUSTRATED

Sheridan

The Data: Two problems illustrated…

Missing Predictors

How far away does the student live?

Data Partition

User Defined | Output | Notes

				Format	Label
			ype	Format	Label
			har	$23.	Unique
			har	$3.	Fees Paid
STD_ID	rejected	nominal	char	$9.	STD ID
ADMIT_TERM	rejected	unary	char	$4.	Admit Term
TERM_DESCR	rejected	unary	char	$9.	Term Descr
APPL_DT	rejected	nominal	char	$10.	Appl Dt
APPL_NBR	rejected	nominal	char	$8.	Appl Nbr
PROG_NBR	input	nominal	char	$1.	Prog Nbr
SEQUENCE	rejected	nominal	char	$1.	Sequence
ACAD_PLAN	rejected	nominal	char	$5.	Acad Plan
PLAN_DESCR	input	nominal	char	$30.	Plan Descr
CAMPUS	input	nominal	char	$3.	Campus
ACAD_LOAD	input	binary	char	$1.	Acad Load
RESIDENCY	input	nominal	char	$4.	Residency
RESD_DESCR	rejected	nominal	char	$15.	Resd Descr
PROG_STATUS	input	nominal	char	$2.	Prog Status
PROG_ACTN	rejected	binary	char	$4.	Prog Actn
PROG_ACTN_DESCR	input				
ACTION_RSN	rejected				
ACTN_RSN_DESCR	input				
ACTION_DT	rejected				
DAYS_AFTER_JAN_1	input				
MONTH_OF_OFFER	input				

Almost the same thing (colinearity)

These can confuse the result.

CURRENT USES AT SHERIDAN

- **Admissions**
  - Determine 1st semester retention factors for current students, then…
  - Fine-tune admissions requirements within Ministry rules & guidelines, then…
  - Assess opportunities for remedial programs based on success factors
- **Benefits**: Admit students most ready and able.

  Provide targeted remedial activity.

CURRENT USES AT SHERIDAN

Student Services

- Assess value of Peer Mentoring to success of recipients and mentors
- Assess value of Orientation programs, fine tune offerings
- Identify retention problems – admission requirements, program problems, killer courses, ???

AND GROWING…

- **Develop an overall student success model**
- **Use the findings to generate more success**

1st Term retention

Program Completion

Employability

Career Success

College/Program Reputation

OPPORTUNITIES

- Acquiring vs. retaining a student
- Short-run planning by Registrar
- Long-run facilities planning
- Direct Mail/Telephone
- Alumni fund raising, post-grad recruitment

OPPORTUNITIES

- **Estimate students acquired…**
  - **Per marketing $ (general, direct, recruitment)**
  - **Per unit change in admission requirements**
- **Estimate students retained…**
  - **Per $ spent on student service A, B, C…**
  - **Per $ spent on faculty training**
  - **Per unit change in avg. student quality, etc.**
- **Optimal allocation of limited resources**

Questions


Thank you!