

Optimize ROI with Lifetime Value Modeling

March 25, 2005

BASUG

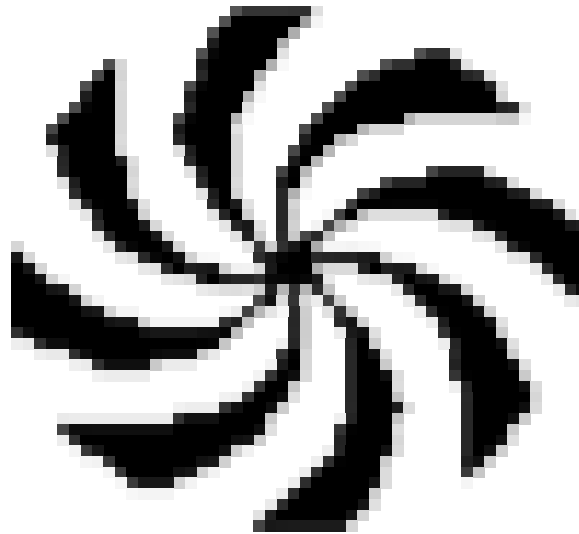
C. Olivia Parr Rud

Agenda

- What is Lifetime Value
 - Advantages
 - Challenges
- Components of LTV
- Sources of Modeling Data
- Case Study
 - Objective Function
- Getting, Exploring and Cleaning the Data
- Choosing the Methodology
- Preparing the Variables for Modeling
- Processing & Validating the Model
 - Decile Analysis
 - Bootstrapping
- Implementing the Model
- Evaluating the Model



Do we think in patterns?



Math Test



30

+ 1000

1030

+ 1000

2030

+ 20

2050

+ 1000

3050

+ 40

3090

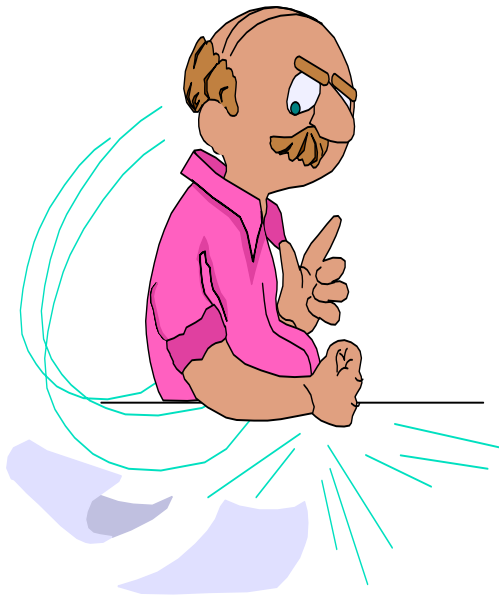
+ 1000

4090

+ 10

Is that your final answer???

4,100



Word Quiz

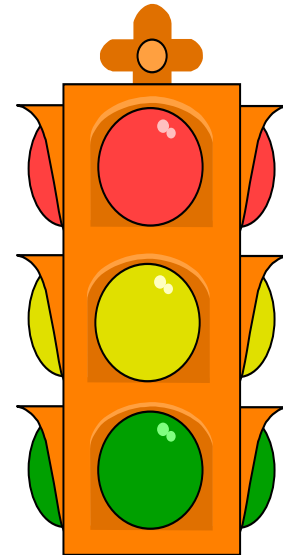




TOP

What do you do at a green light?

GO

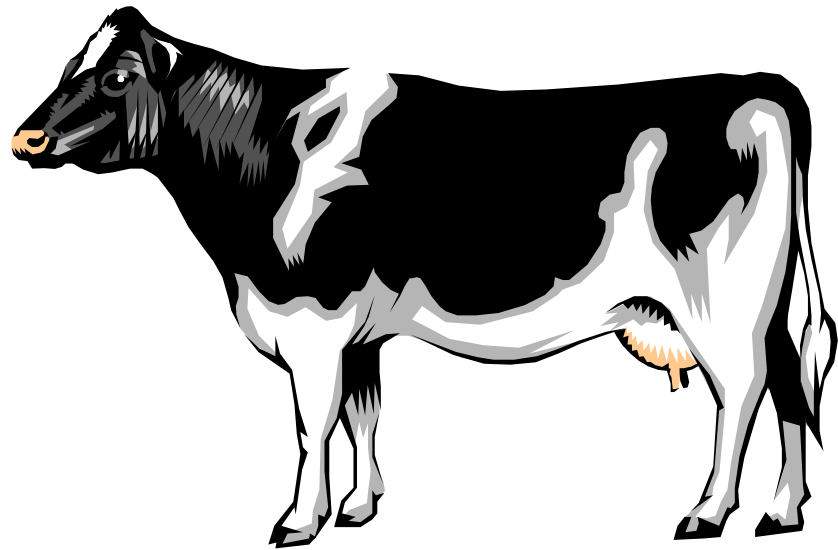




SILK

What does a cow drink?

Water

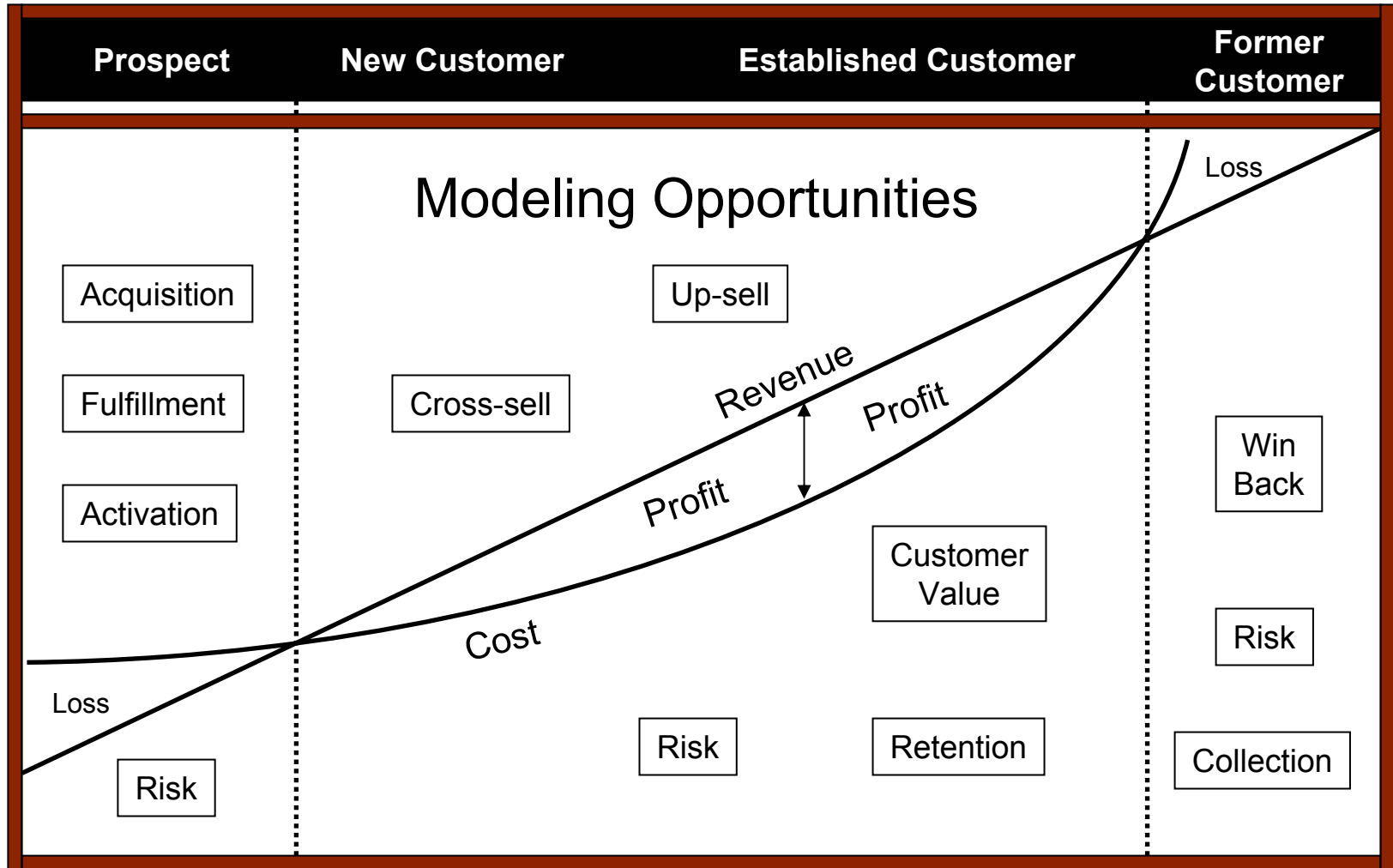


What Is Lifetime Value?

- Present value of a prospect or customer
- Future profits discounted to today's dollars
- Calculation varies by industry and application
- Heavily influenced by the number of contacts
 - Too few can cause LTV to decrease
 - Too many can also cause LTV to decrease
- Often predicted using a series of models

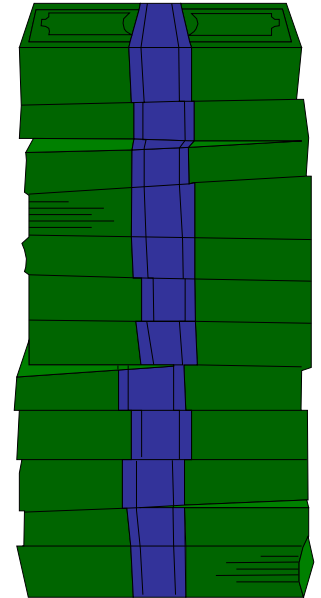
A method to optimize long-term profits

The Customer Lifecycle



What Are Some Advantages of LTV?

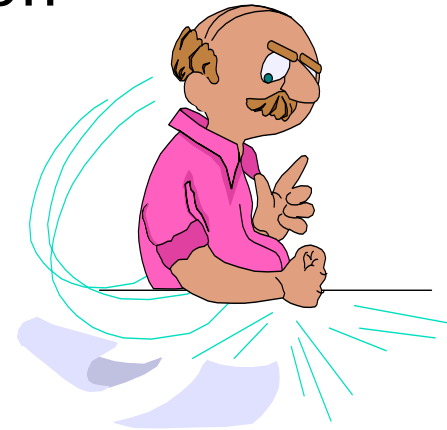
- May increase overall value of the company
- Can be linked directly to marketing goals such as sales targets and customer retention
- Requires company to take the long view
- May account for differences in risk level and timing of customer profit streams



➤ William Burns, Ph.D., *Data Mining Cookbook*, Wiley (2001)

What Are Some Challenges of LTV?

- May compete with short-term goals related to stock value
- Requires integrated data on customers and/or prospects
- Complexity can impede validation



Components of LTV

- Duration
 - Expected length of customer relationship
- Time Period
 - Length of incremental relationship
- Revenue
 - Income from sale of product or service
- Costs
 - Marketing expense or direct cost of product

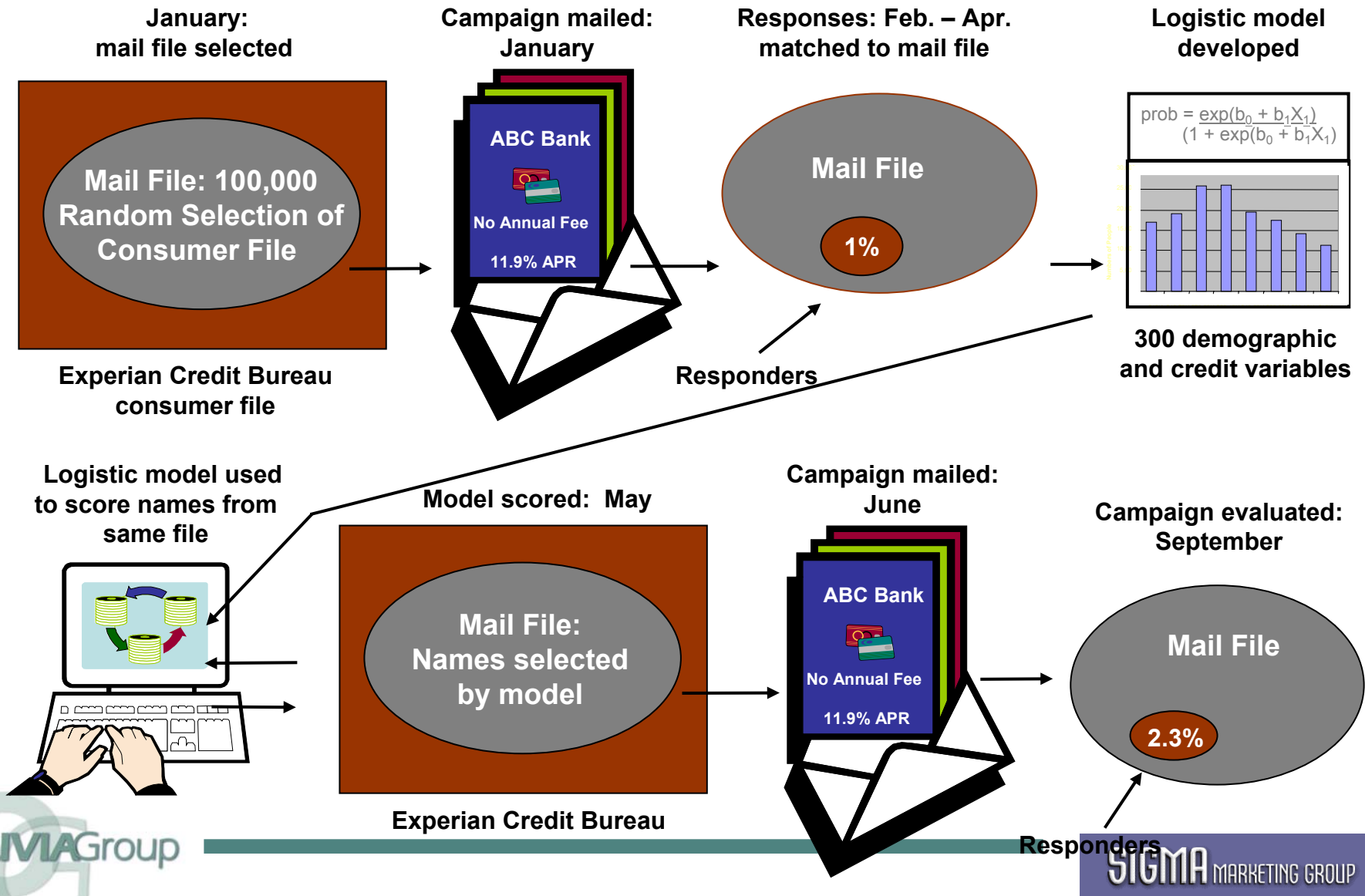
Components of LTV?

- Discount Rate
 - Adjustment to convert future dollars to today's value
- Renewal Rate
 - Probability of renewal or retention
- Referral Rate
 - Incremental revenue generated
- Risk Factor
 - Potential losses related to risk of claim, default, or fraud

Data for Modeling



Data for Modeling Response and Activation



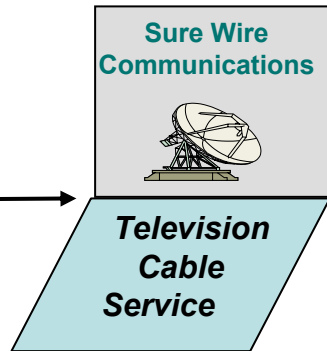
Model to Predict Cross-sell

Phone survey to sample
of customer database

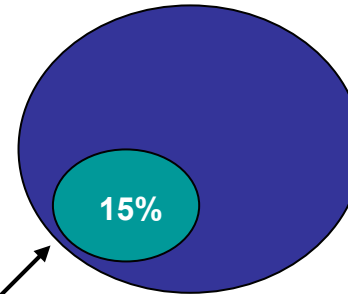


Sure Wire
Customer File

Interest in Cable
Service

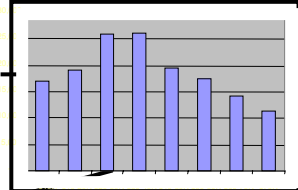


Say "yes" to
cable television



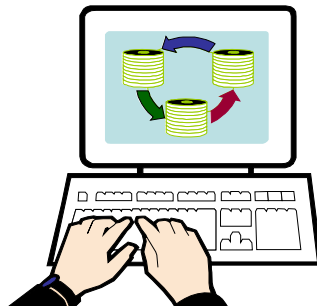
Logistic model
developed

$$\text{prob} = \frac{\exp(b_0 + b_1 X_1)}{1 + \exp(b_0 + b_1 X_1)}$$



Total list demographics
and customer behavior

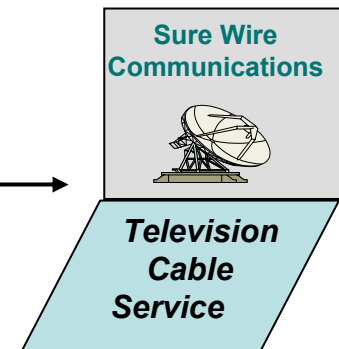
Logistic model used
to score and rank
customer file



Decile Analysis

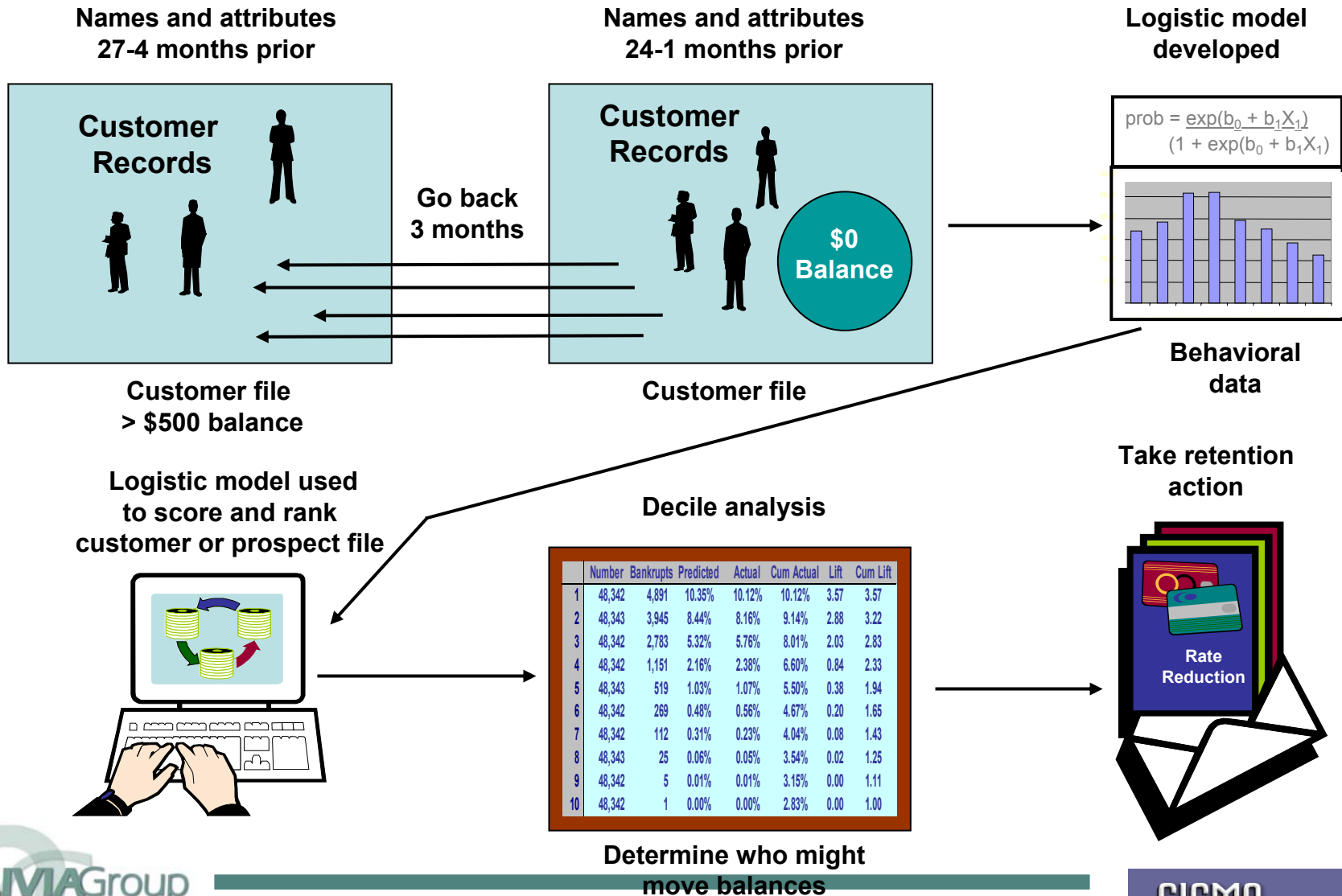
	Number	Bankrupts	Predicted	Actual	Cum Actual	Lift	Cum Lift
1	48,342	4,891	10.35%	10.12%	10.12%	3.57	3.57
2	48,343	3,945	8.44%	8.16%	9.14%	2.88	3.22
3	48,342	2,783	5.32%	5.76%	8.01%	2.03	2.83
4	48,342	1,151	2.16%	2.38%	6.60%	0.84	2.33
5	48,343	519	1.03%	1.07%	5.50%	0.38	1.94
6	48,342	269	0.48%	0.56%	4.67%	0.20	1.65
7	48,342	112	0.31%	0.23%	4.04%	0.08	1.43
8	48,343	25	0.06%	0.05%	3.54%	0.02	1.25
9	48,342	5	0.01%	0.01%	3.15%	0.00	1.11
10	48,342	1	0.00%	0.00%	2.83%	0.00	1.00

Mail Campaign

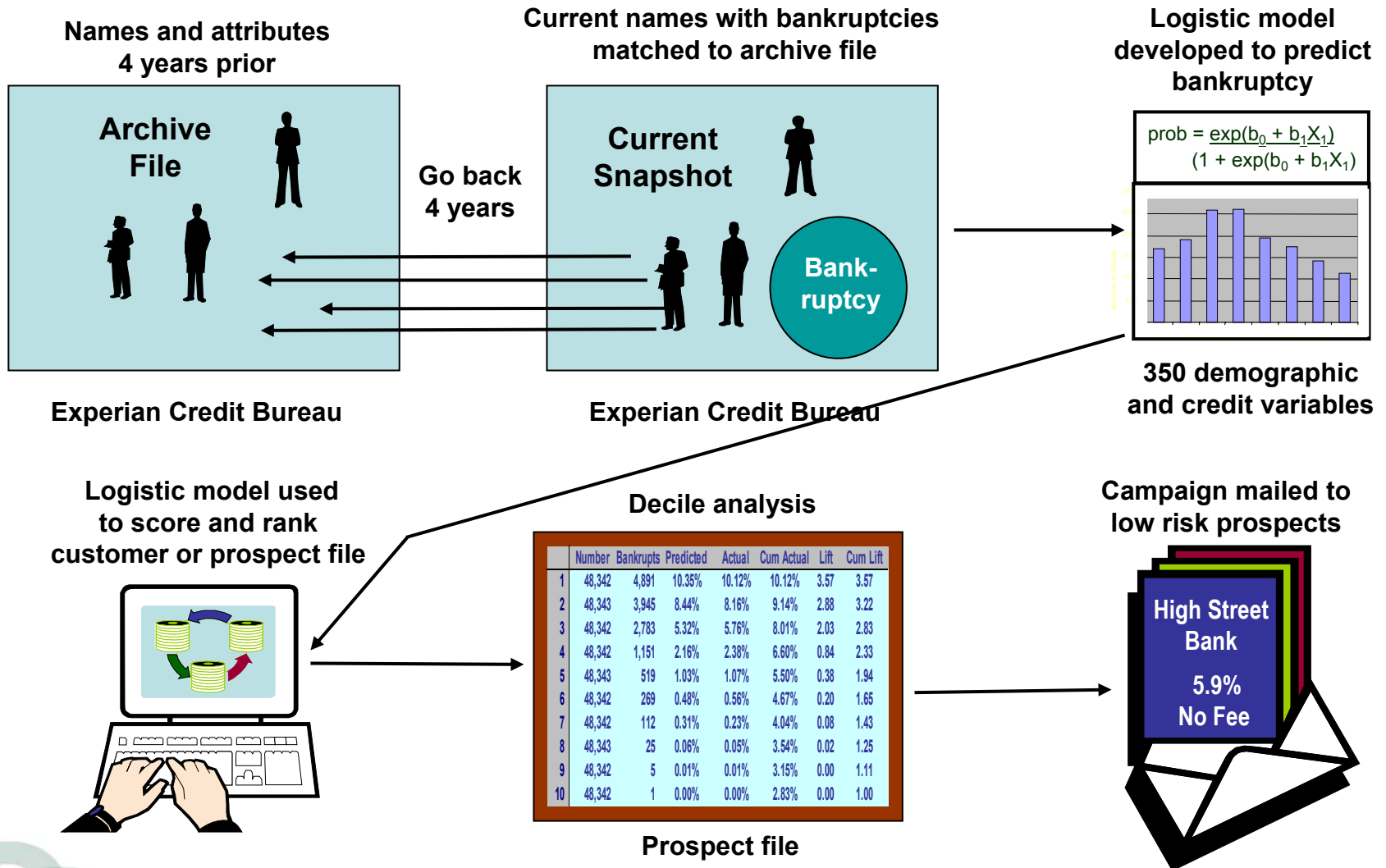


Names selected from
Sure Wire Customer File

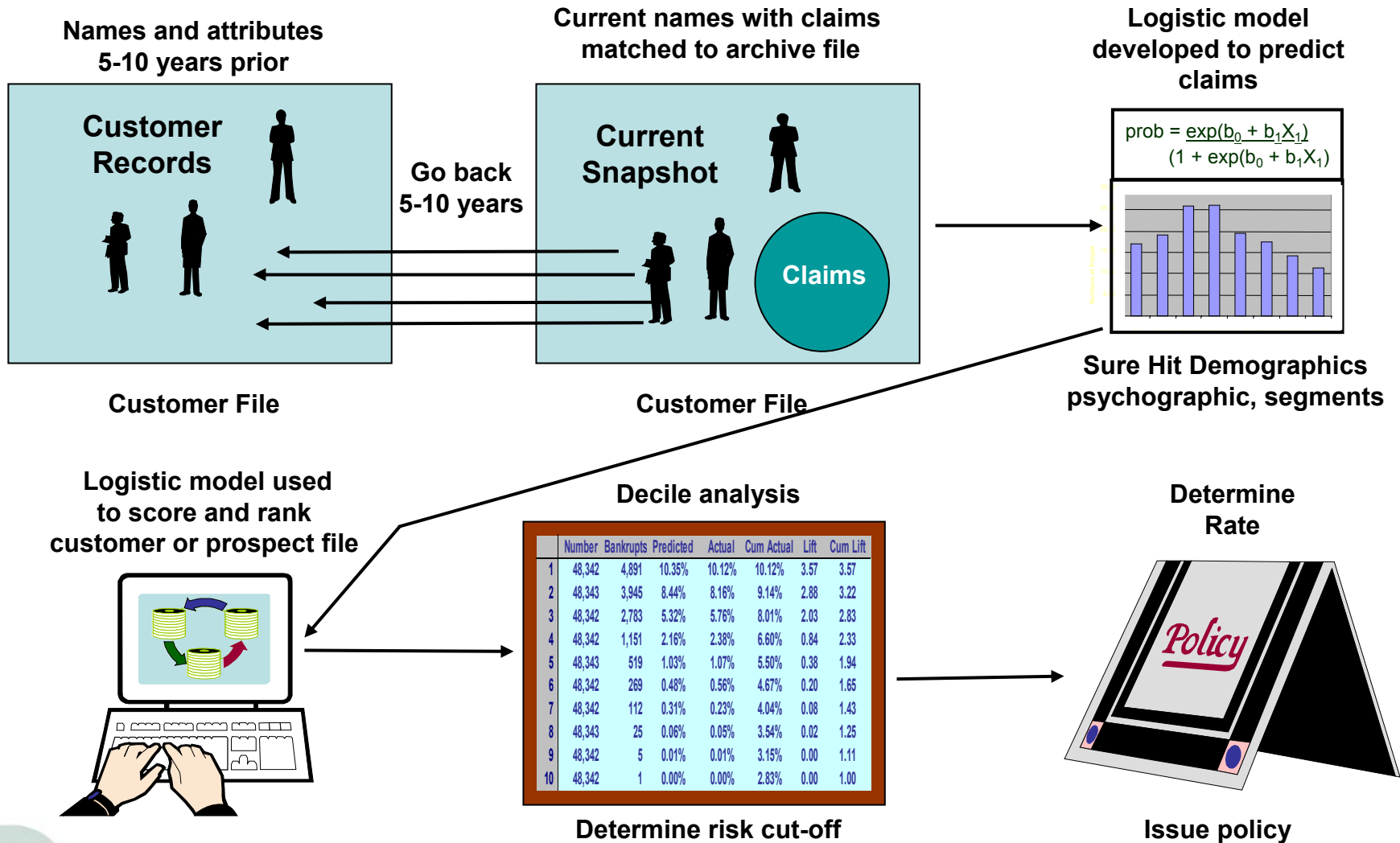
Model to Predict Attrition/ Churn



Risk Model to Predict Bankruptcy

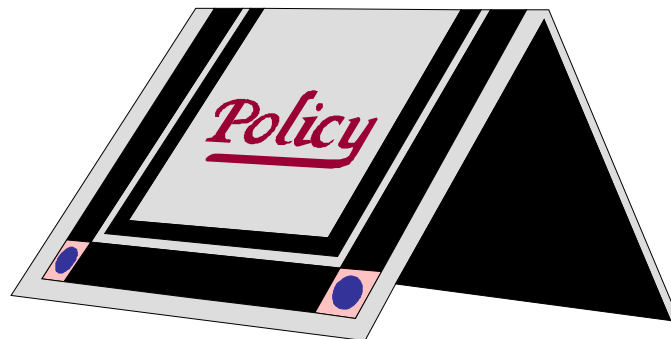


Risk Model to Predict Insurance Claims



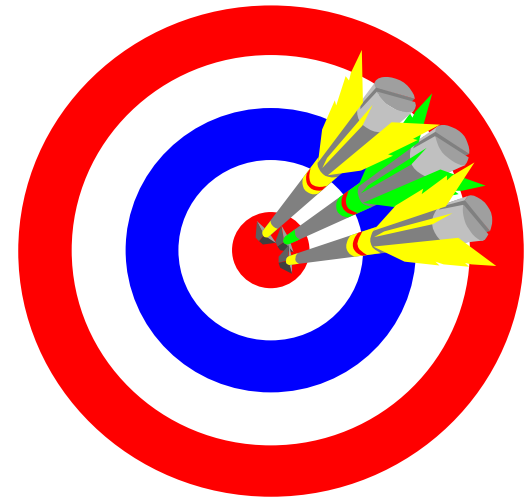
Lifetime Value for Direct Mail Insurance

Case Study



The Objective Function

- What do we want to measure?
LTV = Lifetime Value –
potential customer profitability
for a five year period
- Goal: To target our solicitations to
those prospects who are most
likely to be profitable.



Defining Lifetime Value

$$\begin{aligned} \text{LTV} = & \text{P(Active)} \\ & \times \text{Risk Index} \\ & \times (\text{Product Profitability} \\ & \quad + \text{Cross-Sell/Up-Sell}) \\ & \times \text{Retention Index} \\ & - \text{Marketing Expense} \end{aligned}$$

- LTV is measured at a prospect level in dollars and cents.
- Product profitability and cross-sell/up-sell revenues form the base.
- The probability of becoming active as well as the risk and retention indices are multiplied times the revenue base.
- The marketing expense is subtracted from the total.

Components of LTV

LTV = P(Active)

× Risk Index

× (Product Profitability

+ Cross-Sell/Up-Sell)

× Retention Index

– Marketing Expense

	<i>MALE</i>				<i>FEMALE</i>			
<i>Age</i>	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>
<i>< 40</i>	<i>1.22</i>	<i>1.15</i>	<i>1.18</i>	<i>1.10</i>	<i>1.36</i>	<i>1.29</i>	<i>1.21</i>	<i>1.17</i>
<i>40-49</i>	<i>1.12</i>	<i>1.01</i>	<i>1.08</i>	<i>1.02</i>	<i>1.25</i>	<i>1.18</i>	<i>1.13</i>	<i>1.09</i>
<i>50-59</i>	<i>0.98</i>	<i>0.92</i>	<i>0.90</i>	<i>0.85</i>	<i>1.13</i>	<i>1.08</i>	<i>1.10</i>	<i>1.01</i>
<i>60+</i>	<i>0.85</i>	<i>0.74</i>	<i>0.80</i>	<i>0.79</i>	<i>1.03</i>	<i>0.98</i>	<i>0.93</i>	<i>0.88</i>

Data for Analysis

- Previous campaign mail tape (6 months prior) with response and premium payment appended.
- Cross sell amount from customer profit model.
 - Model score that predicts additional profit based on similar first product.
- Product profitability from product manager (5-year discounted): \$553.
- Lapse indicator adjustment (1.14 – autopay, .86 – other)
- Marketing expense: \$.78 per piece.
- Risk matrix from actuarial.



Preparing the Data

- Sample the 'non active' population and determine weights
- Assess data quality
- Partition data
 - Development dataset
 - Validation dataset



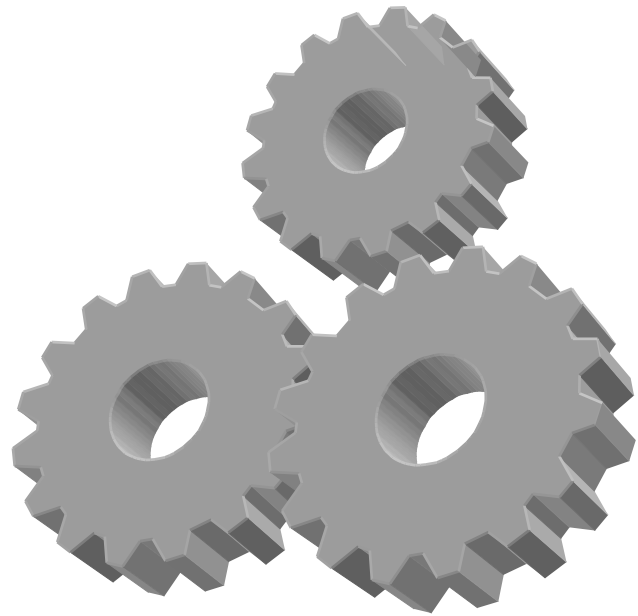
Sampling and Assigning Weights

- Take a random sample of the non-responders and non-paid responders.

	<i>Campaign</i>	<i>Sample</i>	<i>Weight</i>
<i>Non Resp/Non Pd Resp</i>	929,075	37,163	25
<i>Active Accounts</i>	37,781	37,781	1
<i>Total</i>	966,856	74,944	

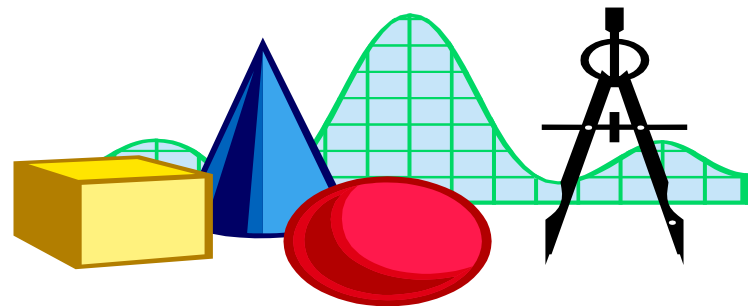
Preparing the Dependent Variable

- Assign dependent variable
- If first-premium > 0
then ACTIVE = 1;
else ACTIVE = 0;



Preparing the Data

- Leverage and replace missing values
- Handle outliers
- Transform or dichotomize continuous variables
- Collapse format categorical variables



Preparing the Data: Univariate analysis

Variable=INCOME (000)

Weight= SMP-WGT

				<u>Extremes</u>			
<u>Moments</u>				Lowest	Obs	Highest	Obs
N	74915	Sum Weights	966827	0	(73850)	206	(53032)
Mean	60.6484	Sum	58636514	0	(73447)	236	(34395)
Std Dev	9.84982	Variance	9969.987	0	(72867)	253	(48823)
Skewness	.	Kurtosis	.	0	(72409)	256	(74504)
USS	4.30E+09	CSS	7.47E+08	0	(71755)	900	(59604)
CV	164.6372	Std Mean	0.101548				
Missing Value				.			
Count				29			
% Count/Nobs				0.04			

Preparing the Data:

Handling missing values and outliers



- Missing values must be managed or replaced.
- Zeroes can represent missing values – must consider data and business case.
- If number of missing values occurs for a large portion of the dataset and/or numerous variables, consider building separate models.
- Outliers may be treated like missing values.
- Substitute single value: mean, median, mode.
 - advantages and disadvantages
- Substitute derived values: result of cluster analysis.
- Substitute value derived from regression.

Preparing the Data:

Missing value substitution

- Cluster analysis groups variables with like characteristics and find the mean within the cluster.

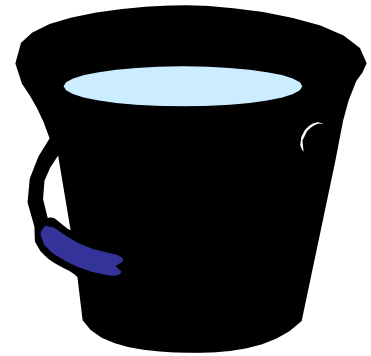
	<i>MALE</i>				<i>FEMALE</i>			
	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>
<i>< 40</i>	55	58	54	45	35	52	46	31
<i>40-49</i>	57	60	58	55	40	54	49	37
<i>50-59</i>	61	63	60	58	46	55	51	42
<i>60+</i>	58	60	61	55	44	46	36	27

Preparing the Variables: Continuous

- Transform variables to improve linear relationship
- Univariate logistic: throw out any variables with a chi-square p-value $> .5$ or use it to select the best 100 prospect variables
- Square, cube, sq root, cube root, log, inverse, etc.
- Create bins or buckets to handle discrete patterns
- Quick method – create every transformation for every continuous variable; run stepwise logistic; let the modeling process select the best variable transformations

Binning / Bucketing / Segmentation

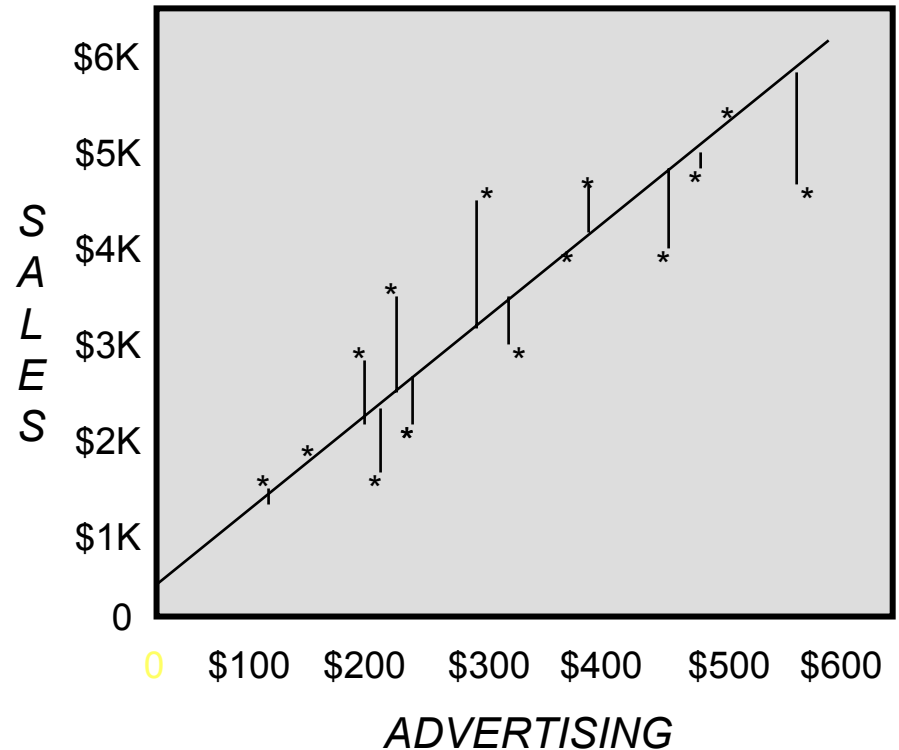
- Useful to create segments for obviously discrete groups
- Capture power of variable by using categorical form
- Why is it necessary?
 - Look at example using linear regression



NOTE: The three terms are often used interchangeably.

Simple Linear Regression

➤ Very linear relationship



➤ $\text{Sales} = 866.59 + 7.81 * \text{Advertising}$

Binning the Predictor

➤ Relationship

2K: advertising < 300

2.5K: 300 ≤ advertising < 400

2.5K: 400 ≤ advertising

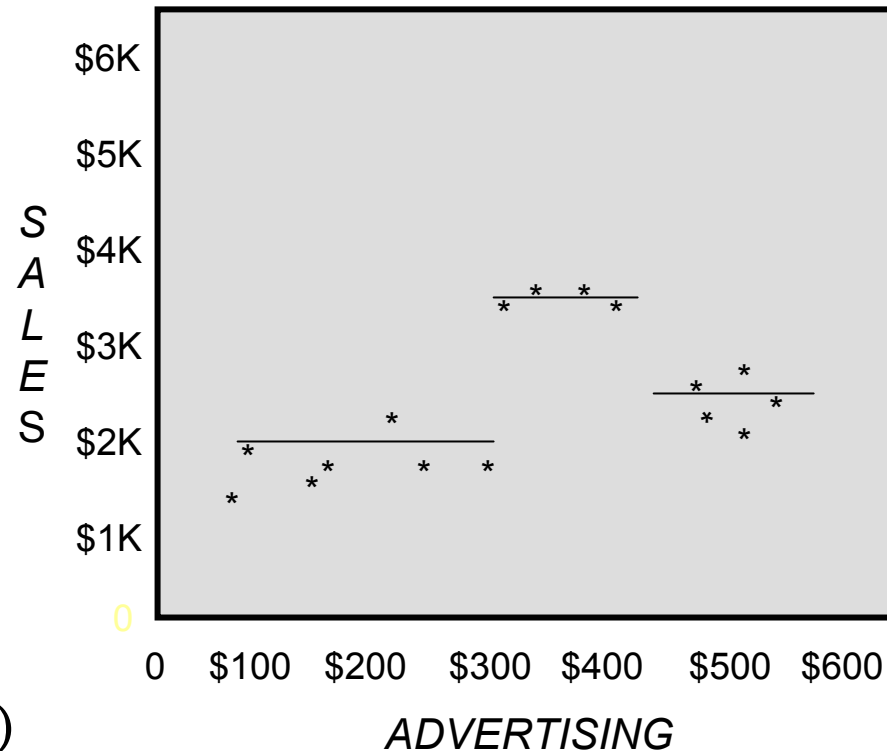
➤ Suggested Variables

B1 = (advertising < 300)

B2 = (300 ≤ advertising < 400)

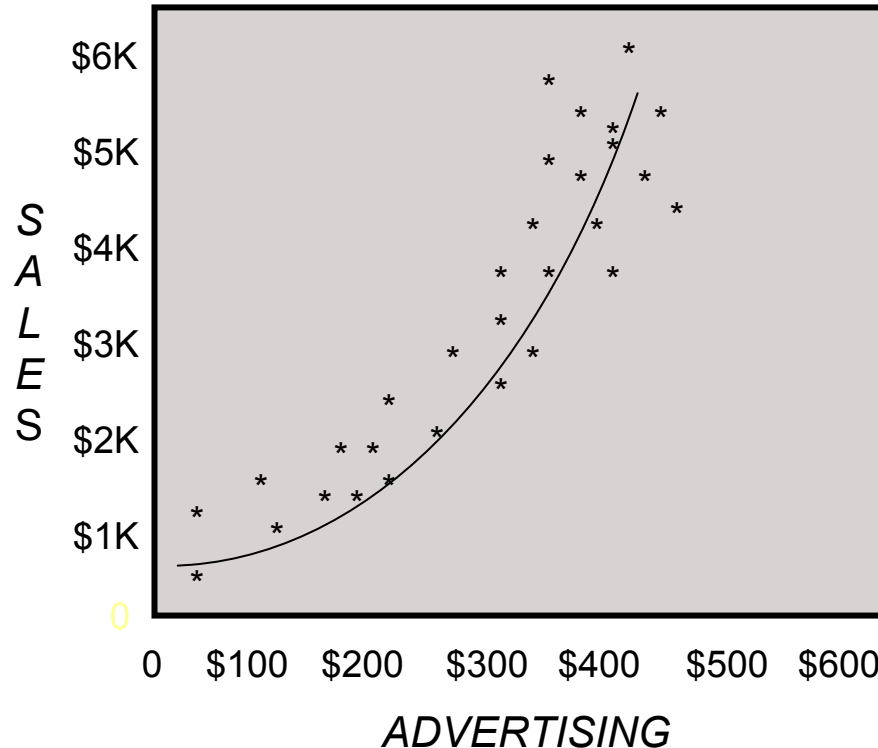
B3 = (400 ≤ advertising)

➤ Value is 1 if logic is true, otherwise 0.



Linear Transformation

➤ If we square the value for Advertising, we get a better fit.



➤
$$\text{Sales} = 749.34 + 0.437 * \text{Advertising}^2$$

Preparing the Variables: Categorical

Frequency					
Col Pct	East	South	Midwest	West	Total
Non Active	232903 96.16	267475 96.76	186805 95.39	241892 96.62	929075 96.09
Active	9298 3.27	8944 2.24	8859 4.43	10680 4.28	37781 3.91
Total	242201 25.05	276419 28.59	195664 20.24	252572 26.12	966856 100.00

Midwest and West have similar performance and can be grouped together

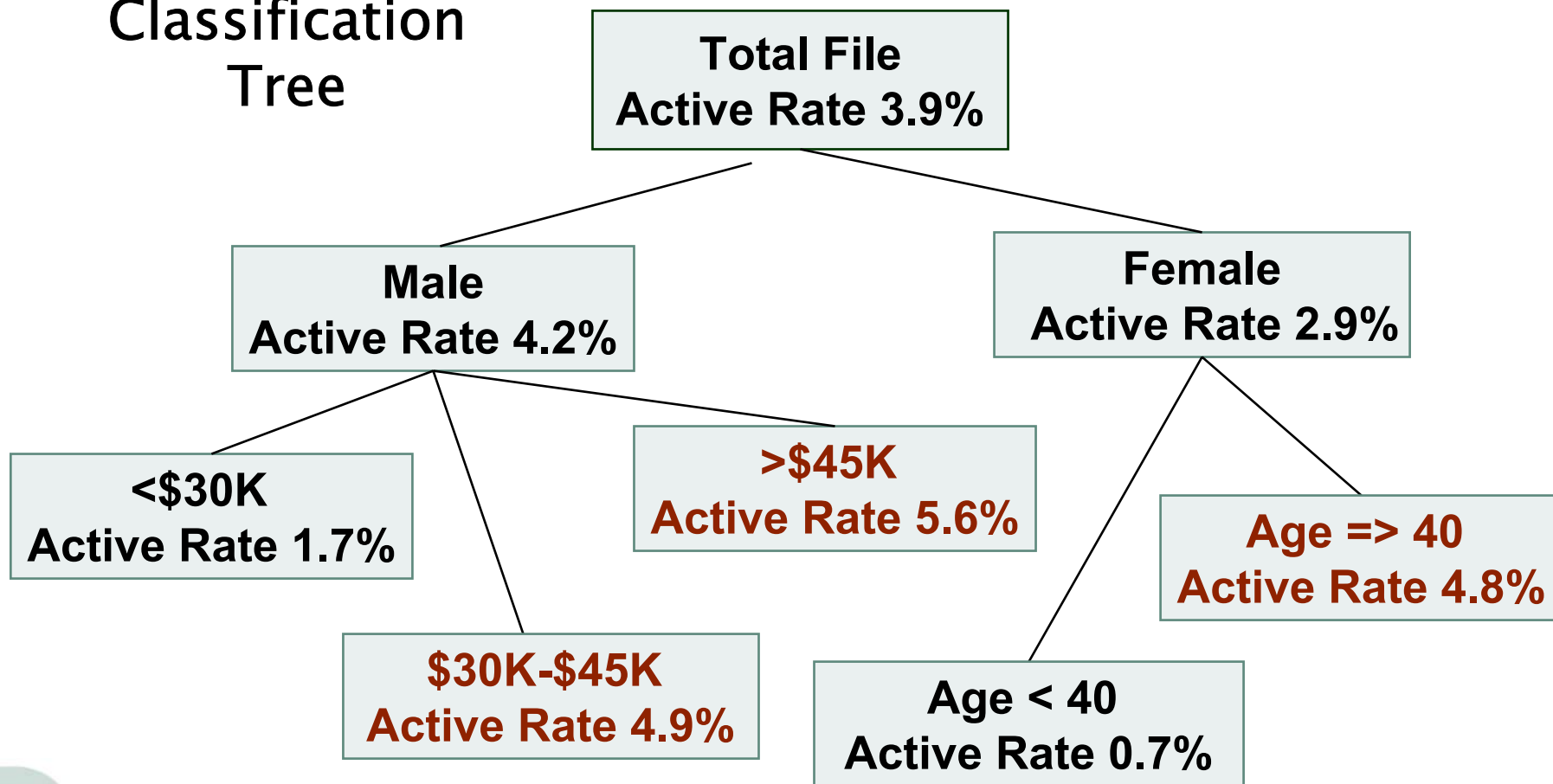
Preparing the Variables: Categorical

- Assign numeric values to categorical variables
- REGION = 'Midwest' and 'West' is the default

Indicator Variables	East	South
East	1	0
South	0	1
Midwest	0	0
West	0	0

Preparing the Variables: Interaction Detection

Classification Tree



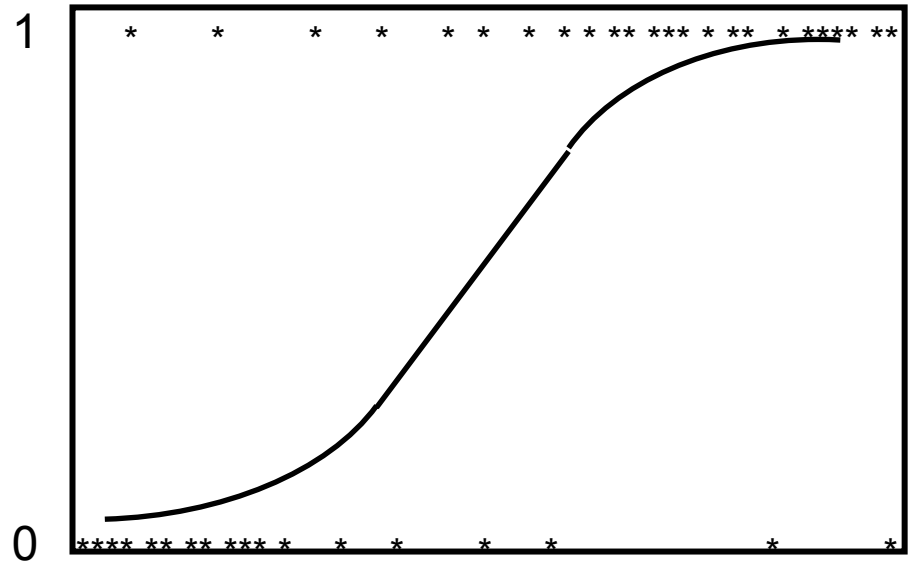
Processing the Model

- Logistic regression – continuous predictors/binary outcome
- Select method – backward, stepwise, best subsets
 - Backward, stepwise – look for best 40.
 - Score selection – get two best models for each number of variables; options selection=score best=2
- Assign weights – sample weight
- Assign p-value – default is .05
- Choose final variables:
AGE, AGE_CUBE , AGE_MISS, MALE, M_INC_LT45, F_AGE_GT40, EAST, SOUTH, INCOME, INC_MISS, LOG_INC, MARRIED, SINGLE, POPDENS, MAIL_ORD



Logistic Regression

- Predicts probability of event occurring using function of linear predictors.
- p = probability of event occurring.
- $p/(1-p)$ is the odds of an event occurring.
- Log of the odds:
 $\log(p/(1-p))$ is linear function of predictors.



Uses sigmoidal function instead of linear function to fit the data.

$$\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Model Output

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi_Square
Intercept	1	0.1526	0.1856	0.6764	0.4108
AGE_CUBE	1	-0.0863	0.0162	152.3548	0.0001
EAST	1	0.6263	0.0363	368.0549	0.0001
LOG_INC	1	0.0194	0.0024	60.164	0.0001
INC_MISS	1	-0.5220	0.0326	251.7456	0.0001
MARRIED	1	0.7758	0.0279	156.2743	0.0001
F_AGE_GT40	1	0.4562	0.0368	153.4113	0.0001
MAIL_ORD	1	0.5693	0.0270	443.3525	0.0001

$$p(\text{active}) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)})$$

$$p(\text{active}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}{(1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n))}$$

Validating the Model

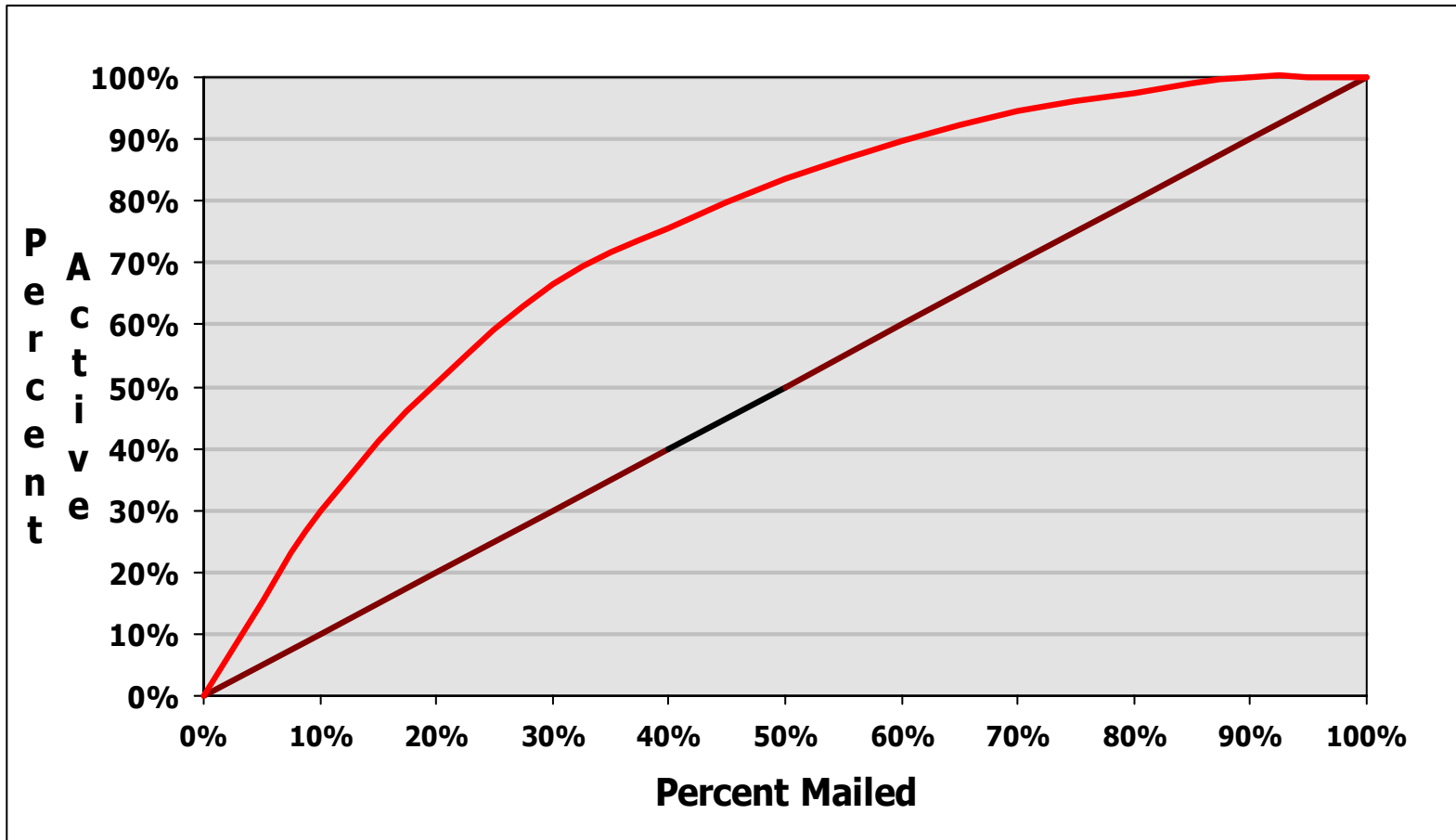
- Score hold out sample and create gains table.
- Whenever possible, score a campaign or file from a different time period.
- Use bootstrapping to create confidence intervals around estimates.

Always validate on full sample of data set.

Gains Table on Validation Data

	Number	Accounts	Predicted	Actual	Cum Actual	Lift	Cum Lift
1	48,342	4,891	10.35%	10.12%	10.12%	3.57	3.57
2	48,343	3,945	8.44%	8.16%	9.14%	2.88	3.22
3	48,342	2,783	5.32%	5.76%	8.01%	2.03	2.83
4	48,342	1,151	2.16%	2.38%	6.60%	0.84	2.33
5	48,343	519	1.03%	1.07%	5.50%	0.38	1.94
6	48,342	269	0.48%	0.56%	4.67%	0.20	1.65
7	48,342	112	0.31%	0.23%	4.04%	0.08	1.43
8	48,343	40	0.06%	0.05%	3.54%	0.02	1.25
9	48,342	15	0.01%	0.01%	3.15%	0.00	1.11
10	48,342	3	0.00%	0.00%	2.83%	0.00	1.00

Gains Chart on Validation Data



Bootstrapping

- Validation technique that uses resampling.
- Provides an empirical estimation based on experience and observation instead of a parametric estimation based on a system or distribution.
- Detects over-fitting by using repeated samples to validate the results of a model.
- Allows for confidence intervals around estimates to aid in campaign performance predictions.



What Is a Bootstrap Sample?

- A sample with n observations pulled from an original sample with n observations.
- Pulled with replacement.
- A single observation from the original sample can appear more than once in a bootstrap sample while other observations may be missing entirely.

Original Sample

1 2 3 4 5 6 7 8 9 10

Bootstrap Sample

1 3 3 4 5 6 7 7 9 10

N = 10

Bootstrap Estimates

- A bootstrap estimate is calculated as follows:
- $$\text{BS Estimate} = 2 * \text{Sample Estimate} - \text{mean}(\text{BSi})$$
- where BSi = the set of bootstrap estimates from the sample.

Confidence Intervals Using Bootstrap Estimates

- To calculate a confidence interval using bootstrap estimates, first derive the standard deviation of the bootstrap samples $\text{Std}(\text{BSi})$.
 - The standard deviation of the bootstrap samples is the sample error of the mean of the bootstrap samples.
- To get a 95% bootstrap confidence interval, use the following formula:
 - $\text{UCI} = \text{BSest} + |Z.025| * \text{Std}(\text{BSi})$
 - $\text{LCI} = \text{BSest} - |Z.025| * \text{Std}(\text{BSi})$

Confidence Intervals for Lift

	Lift	BS Est	Lower CI	Upper CI
1	3.57	3.55	3.41	3.68
2	2.88	2.87	2.68	2.99
3	2.03	2.04	1.91	2.10
4	0.84	0.85	0.72	0.94
5	0.38	0.37	0.29	0.45
6	0.20	0.18	0.15	0.22
7	0.08	0.08	0.04	0.12
8	0.02	0.01	0.00	0.03
9	0.00	0.00	0.00	0.01
10	0.00	0.00	0.00	0.00

Components of LTV

LTV = P(Active)

× Risk Index

× (Product Profitability

+ Cross-Sell/Up-Sell)

× Retention Index

– Marketing Expense

	<i>MALE</i>				<i>FEMALE</i>			
<i>Age</i>	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>	<i>M</i>	<i>S</i>	<i>D</i>	<i>W</i>
<i>< 40</i>	<i>1.22</i>	<i>1.15</i>	<i>1.18</i>	<i>1.10</i>	<i>1.36</i>	<i>1.29</i>	<i>1.21</i>	<i>1.17</i>
<i>40-49</i>	<i>1.12</i>	<i>1.01</i>	<i>1.08</i>	<i>1.02</i>	<i>1.25</i>	<i>1.18</i>	<i>1.13</i>	<i>1.09</i>
<i>50-59</i>	<i>0.98</i>	<i>0.92</i>	<i>0.90</i>	<i>0.85</i>	<i>1.13</i>	<i>1.08</i>	<i>1.10</i>	<i>1.01</i>
<i>60+</i>	<i>0.85</i>	<i>0.74</i>	<i>0.80</i>	<i>0.79</i>	<i>1.03</i>	<i>0.98</i>	<i>0.93</i>	<i>0.88</i>

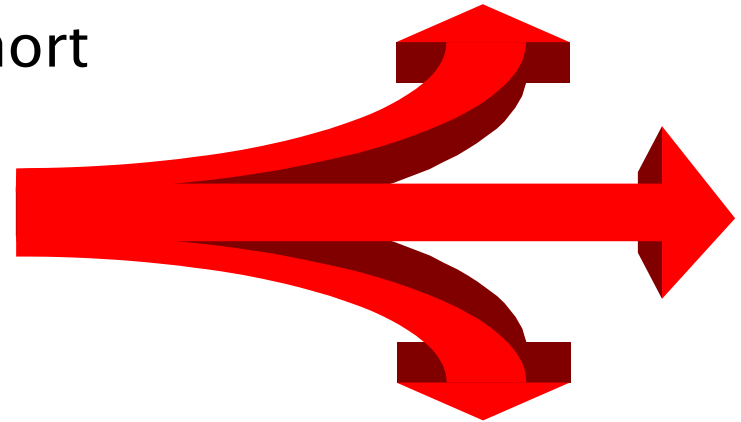
Financial Assessment on New Campaign

	Number	Active Rate	Cross Sell	Risk Index	Lapse Indicator	Product Profitability	Average LTV	Average CUM LTV	Sum Cum LTV
1	96,685	10.36%	\$120	0.94	0.99	\$553	\$64.76	\$64.76	\$6,261,266
2	96,685	8.63%	\$104	0.99	1.00	\$553	\$55.35	\$60.06	\$11,612,984
3	96,685	5.03%	\$105	0.96	0.99	\$553	\$30.99	\$50.37	\$14,609,591
4	96,685	1.94%	\$107	0.93	0.97	\$553	\$11.13	\$40.56	\$15,685,475
5	96,685	0.96%	\$98	1.01	0.99	\$553	\$5.53	\$33.55	\$16,220,346
6	96,685	0.28%	\$101	1.02	1.00	\$553	\$1.09	\$28.14	\$16,325,522
7	96,685	0.11%	\$97	1.03	1.01	\$553	(\$0.04)	\$24.12	\$16,321,311
8	96,685	0.08%	\$98	0.99	1.01	\$553	(\$0.26)	\$21.07	\$16,295,747
9	96,685	0.01%	\$94	1.04	1.02	\$553	(\$0.75)	\$18.64	\$16,223,586
10	96,685	0.00%	\$95	1.09	1.02	\$553	(\$0.78)	\$16.70	\$16,148,199

How many deciles do you mail?

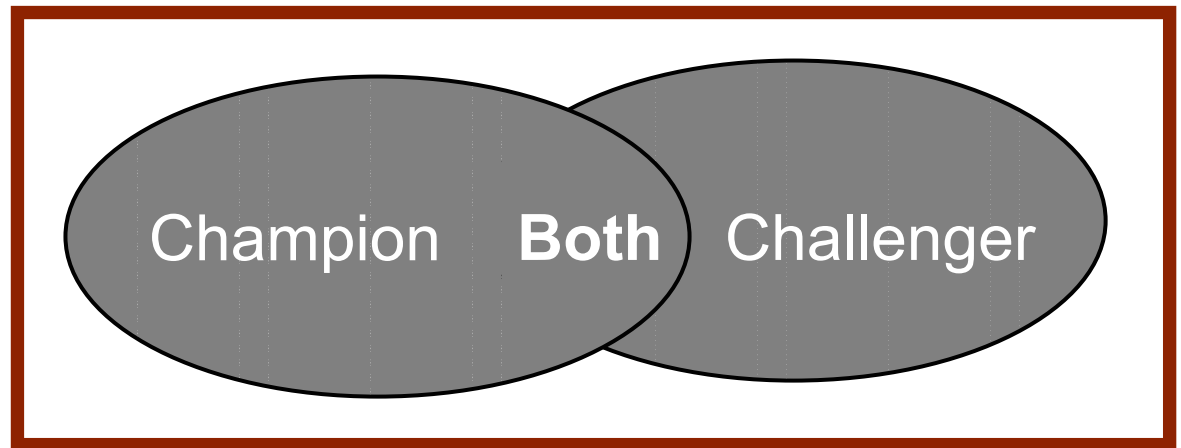
Applying Your Results Correctly

- Is the population to be scored similar to the model development population?
- How long is the evaluation period?
- Will better long term results have a negative effect on short term results?



Implementing the Model

- Select or score file with both models
- Mail all names selected by both models
- Mail random sample of names NOT selected by either model
- Mail all or sample of names preferred by each model



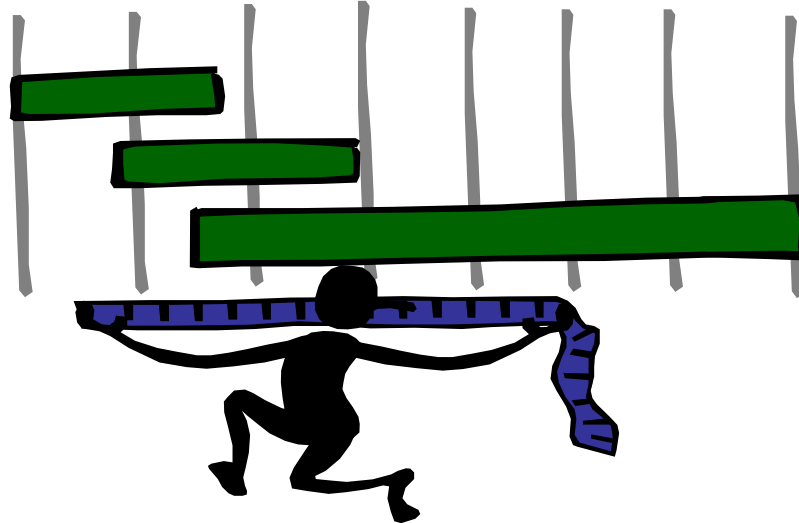
Maintaining the Model

- Expected life of model
- Benchmarking
- Rebuild or refresh?
- Model logs



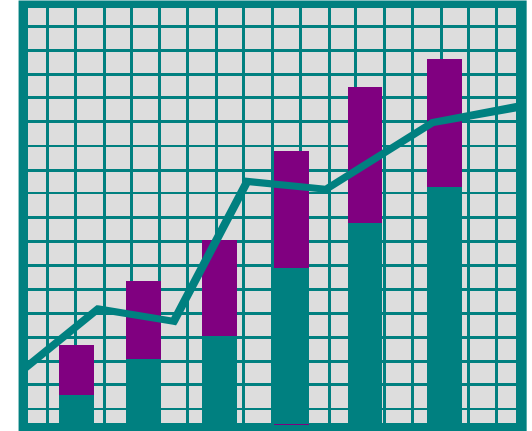
Model Life

- Depends on the target.
 - Response models can be redeveloped within a few months.
 - Risk models need a couple of years to evaluate.
- If the expected life is several years – use benchmarking.



Benchmarking

- Tracks long-term model performance at predefined intervals.
 - Modeling bankruptcy over a three-year period.
 - Take all names that are current for time T .
 - Measure performance in the time period between $T + 6$ to $T + 36$ months.
 - After implementation the performance is measured or benchmarked at each six-month period.
- If the model is not performing as expected
 - Choose whether to continue use, rebuild, or refresh.



Rebuild or Refresh?

- Rebuild – use new data, build new variables, and rework the entire process.
 - Opportunity to introduce new predictive information
 - Current model is very old
 - Strong shifts in the marketplace
 - Product or service offering has changed
- Refresh – keep the current variables and rerun the model on new data.
 - Save time and resources



C. Olivia Parr Rud
olivia@oliviagroup.com
610 918-3801

SIGMA Marketing Group
www.sigmamarketing.com

OLIVIAGroup
www.oliviagroup.com

Author, Data Mining Cookbook (Wiley, 2001)
www.dataminingcookbook.com

Member, Association for Spirit at Work
www.spiritatwork.com

SIGMA Marketing *offering...*

- *Strategic marketing services, knowledge mining, predictive modeling, creative and .*

OLIVIAGroup *offering...*

- *Predictive modeling training, leadership development, change management, communication and creativity training.*

Questions

