New Forms for Business Intelligence

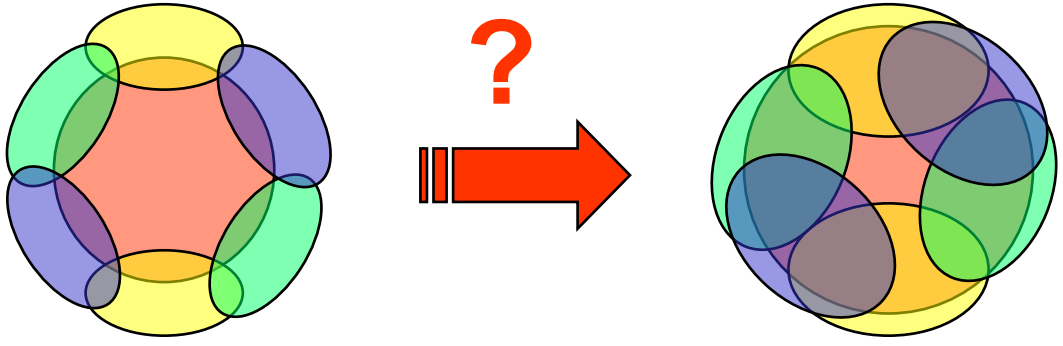
A concrete approach to the use of data to improve and quicken business decisions

Agenda

Module 5: Scoring Models

- Introduction
- Scoring Models & Marketing Campaigns
- The Analytical Process

Business Intelligence & Customer Profiling



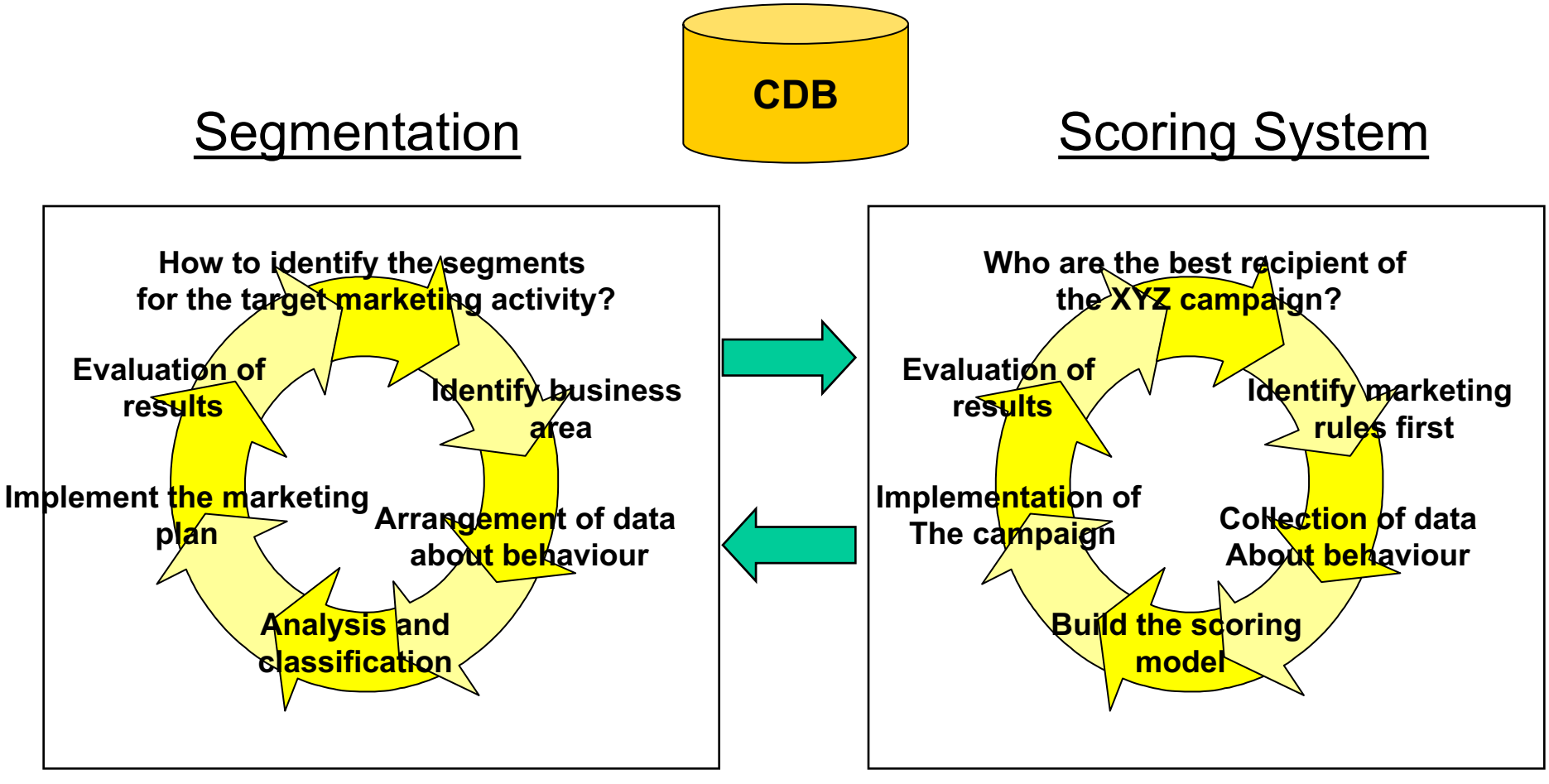
Identification of ‘business’ segments for planning the operations

➔ Segmentation ←

Identification of most likelihood clients to accept a cross/up selling offer

➔ Propensity models←

Goals of a CP project



Predictive Classification

- Analyze the whole population...
  Churn rate is 30%.

- If we divide the population by subcription seniority, we notice that high seniority corresponds to a low churn probability.

Subscription seniority	Churn
<= 10 years	35%
> 10 years	28%

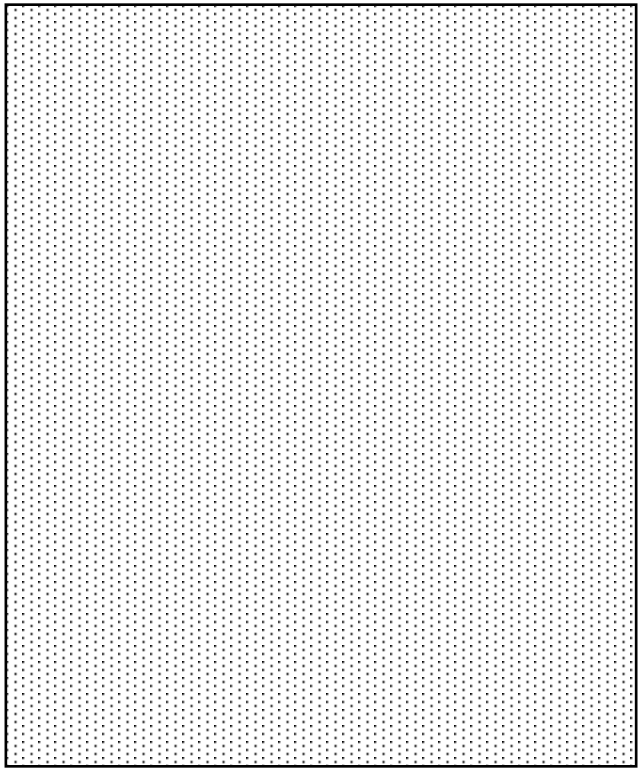
- If we divide the population by subscription seniority and one shot purchases of collateral works to the magazine, differences are even more evident.

Subscription seniority	One shot purchases	Churn
<= 10 years	NO	40%
	YES	30%
> 10 years	NO	38%
	YES	20%

Scoring models

- Support marketing campaign optimization through the identification of customers with more potential.

Customer Database:
sex, age, behaviour



Customer Classification:
redemption wait, sex, …

redemption = 3.9

redemption = 2.8

redemption = 1.6

redemption = 1.0

redemption = 0.4

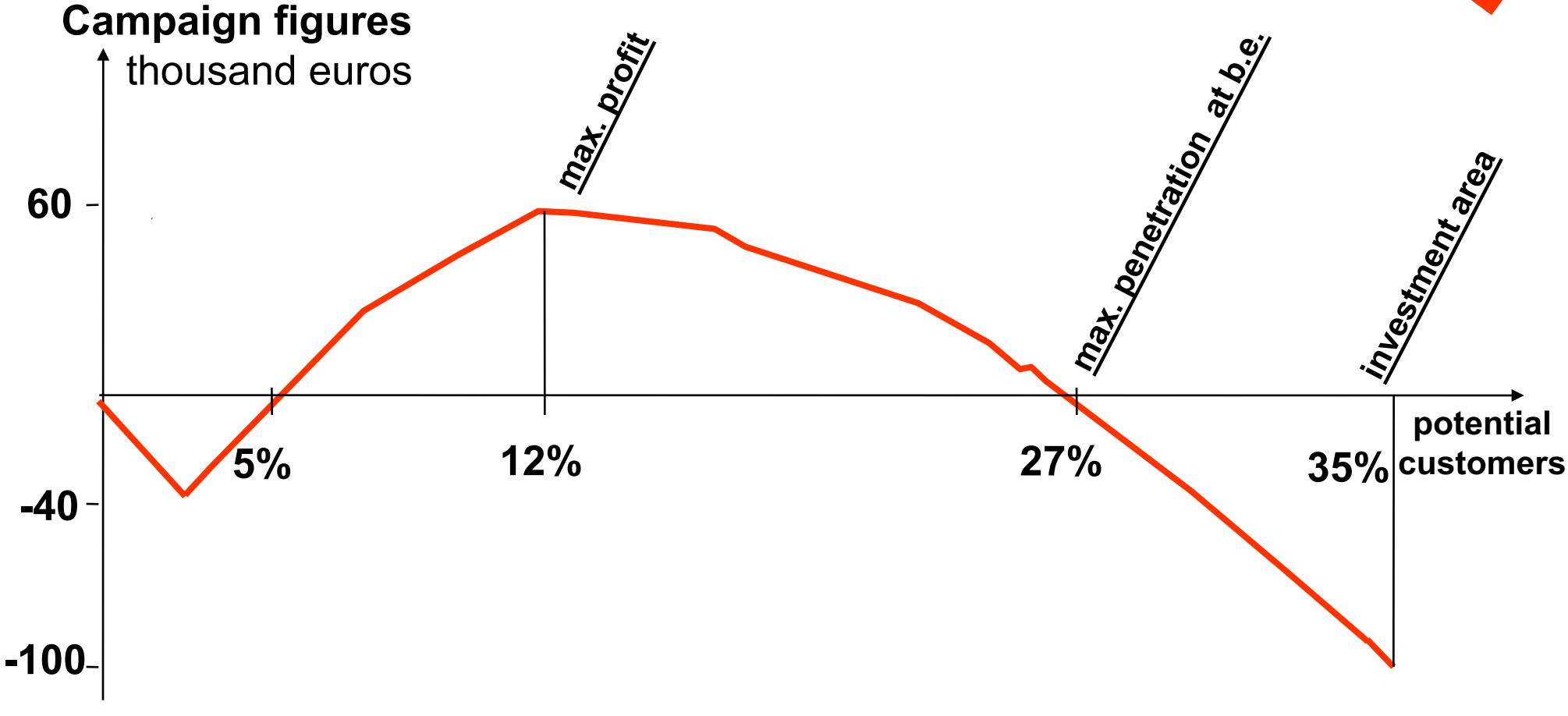
Model usage

example

% Population	Score MIN-MAX	Avg Score	Cum. # Customers	Cum. # Target	Cum. Redemption	Cum. Lift	% Response Captured
5	1:0.1577-0.8055	0.2703	27509	5.00	18.73	8.37	41.83
10	2:0.0835-0.1576	0.1141	55026	10.00	13.51	6.04	60.36
15	3:0.0509-0.0834	0.0655	82537	15.00	10.59	4.73	70.96
20	4:0.033-0.0508	0.0409	110025	20.00	8.74	3.90	78.04
25	5:0.0231-0.0329	0.0276	137442	24.98	7.41	3.31	82.62
30	6:0.0169-0.023	0.0197	165356	30.05	6.42	2.87	86.12
35	7:0.0127-0.0168	0.0146	192621	35.01	5.68	2.54	88.84
40	8:0.0096-0.0126	0.0110	219814	39.95	5.09	2.27	90.88
45	9:0.0072-0.0095	0.0083	247301	44.94	4.62	2.06	92.69
50	10:0.0053-0.0071	0.0061	275370	50.04	4.21	1.88	94.12
55	11:0.0039-0.0052	0.0045	303506	55.16	3.87	1.73	95.36
60	12:0.0029-0.0038	0.0033	329490	59.88	3.60	1.61	96.27
65	13:0.0021-0.0028	0.0024	356163	64.73	3.36	1.50	97.09
70	14:0.0014-0.002	0.0017	386197	70.18	3.12	1.39	97.79
75	15:0.0009-0.0013	0.0011	414908	75.40	2.92	1.31	98.51
80	16:0.0006-0.0008	0.0007	439066	79.79	2.78	1.24	98.97
85	17:0.0003-0.0005	0.0004	474474	86.23	2.58	1.15	99.49
90	18:0.0002-0.0002	0.0002	491813	89.38	2.50	1.11	99.65
95	19:0.0001-0.0001	0.0001	517245	94.00	2.38	1.06	99.86
100	20:0-0	0.0000	550258	100.00	2.24	1.00	100.00

Campaign implementation

example

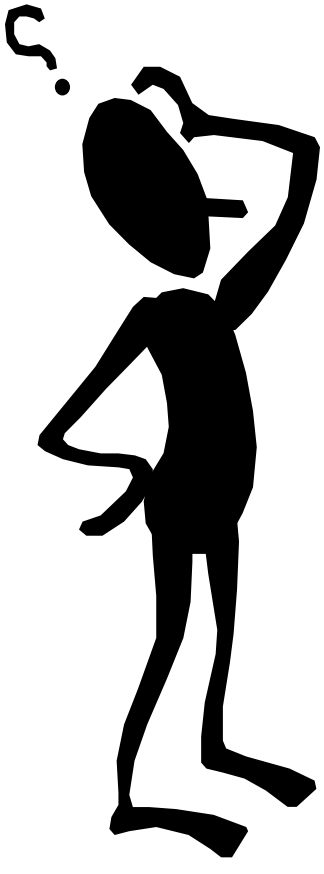


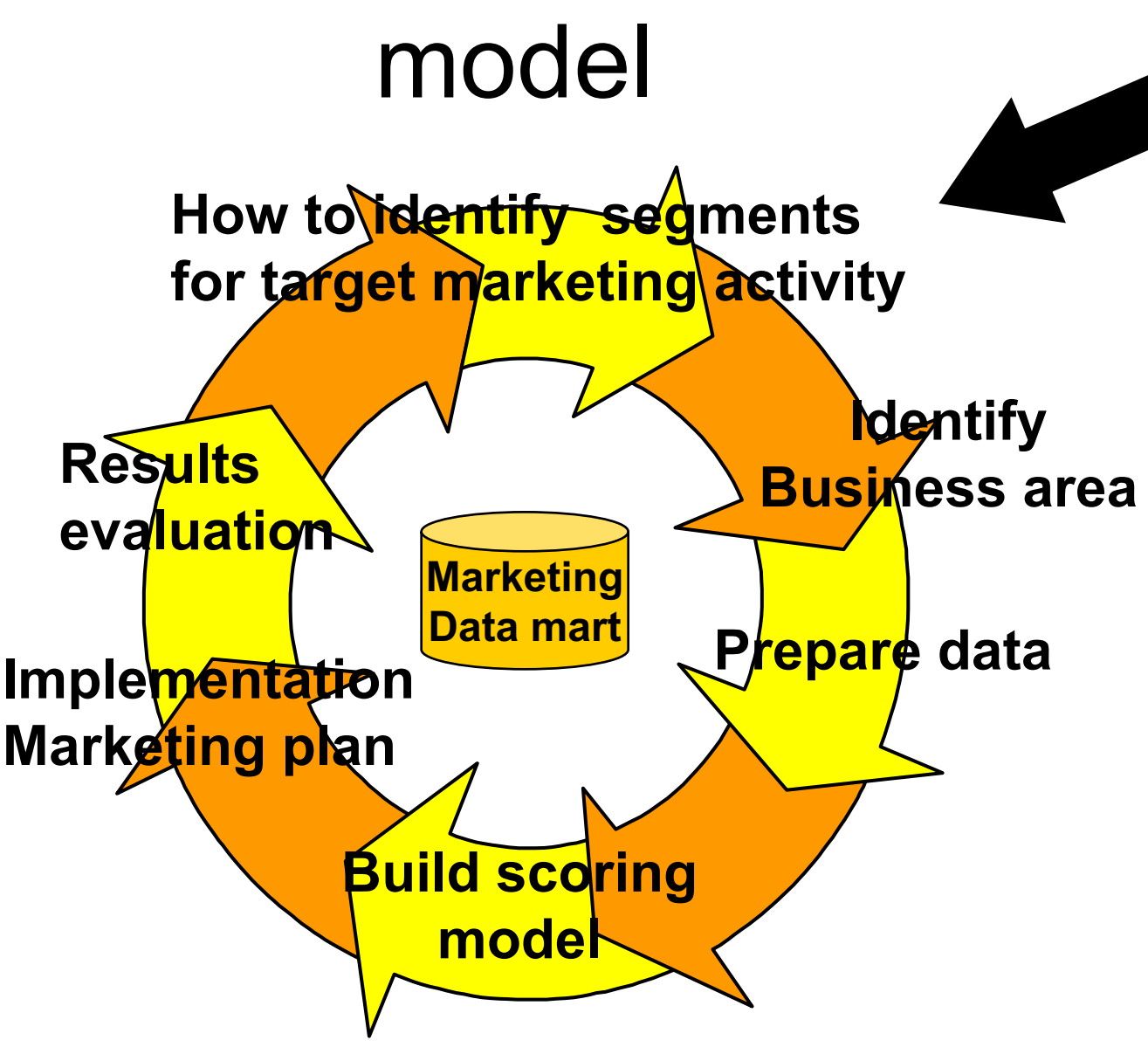
Agenda

Module 5: Scoring Models

- Introduction
- Scoring Models & Marketing Campaigns
- The Analytical Process

Analytical cycle of a propensity model





Project Goals

Propensity models assign a **score** to each client that measures the propensity to buy a certain product.

Building a propensity model for Bond Fund it's important consider the following **marketing goals**:

- Identify clients that are the most inclined to buy Bond Fund with the purpose of offer them the product

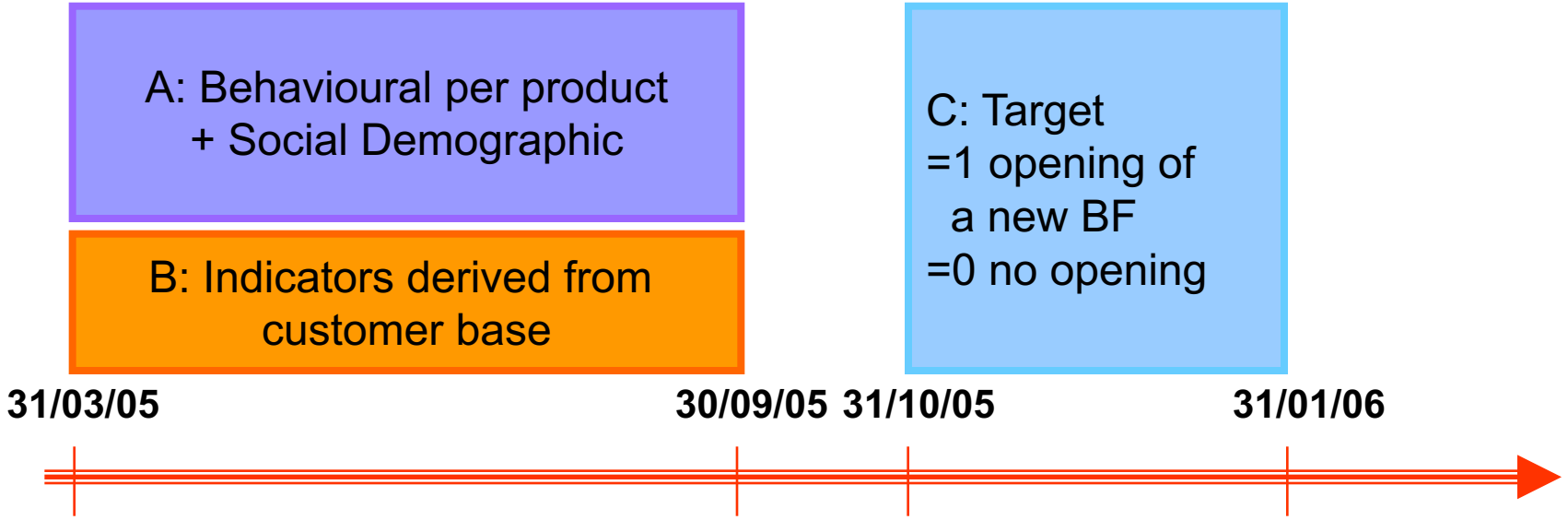
or

- Maximize the number of new Bond Fund sold to clients that have never acquired that product previously (cross-selling), given a fix budget of costs (for example contact not more than 200.000 potential clients)

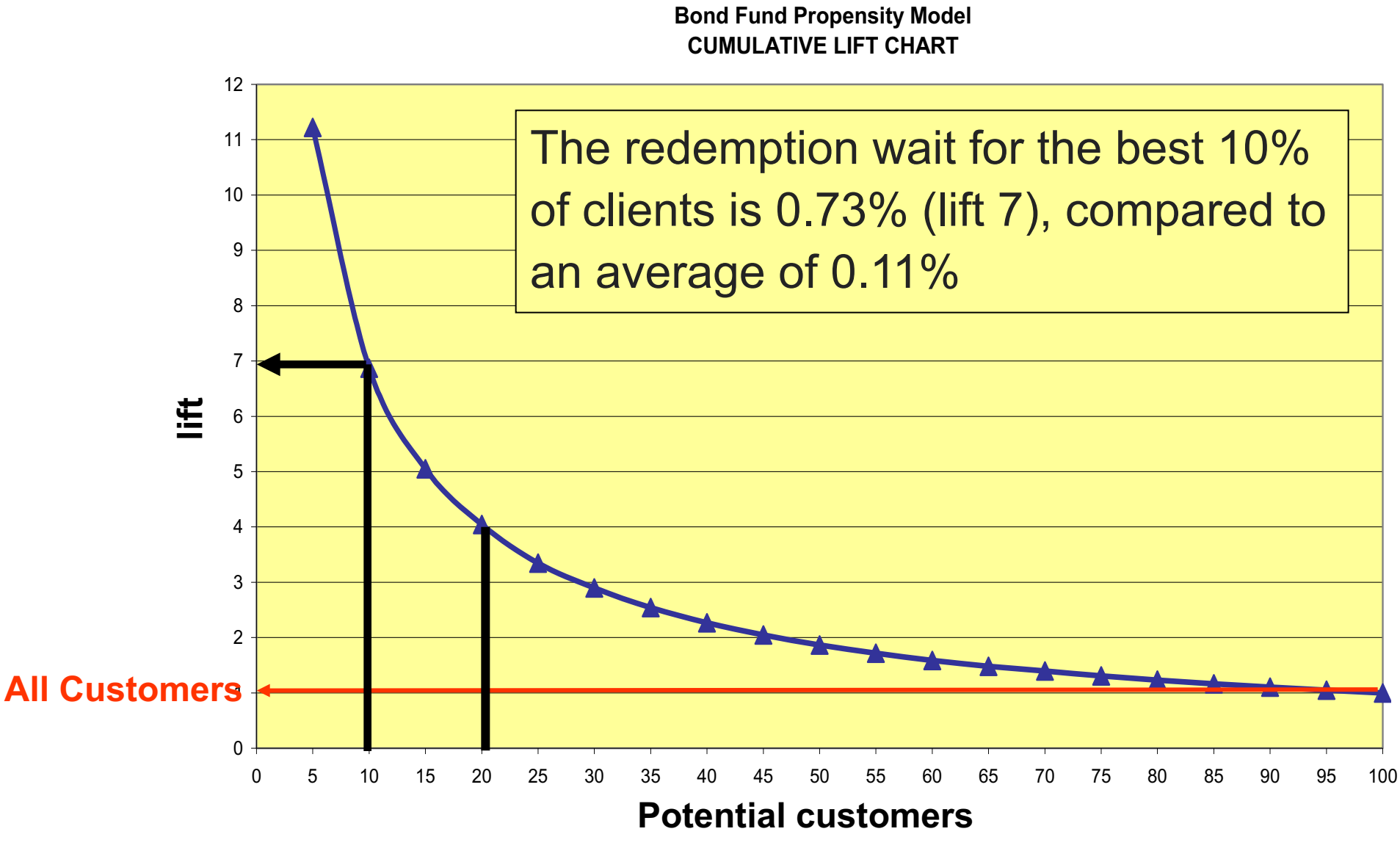
Prepare data

For each client with at least one account open in date 30/09/2005, the CT for the development of the model contains the 3 following sets of variables:

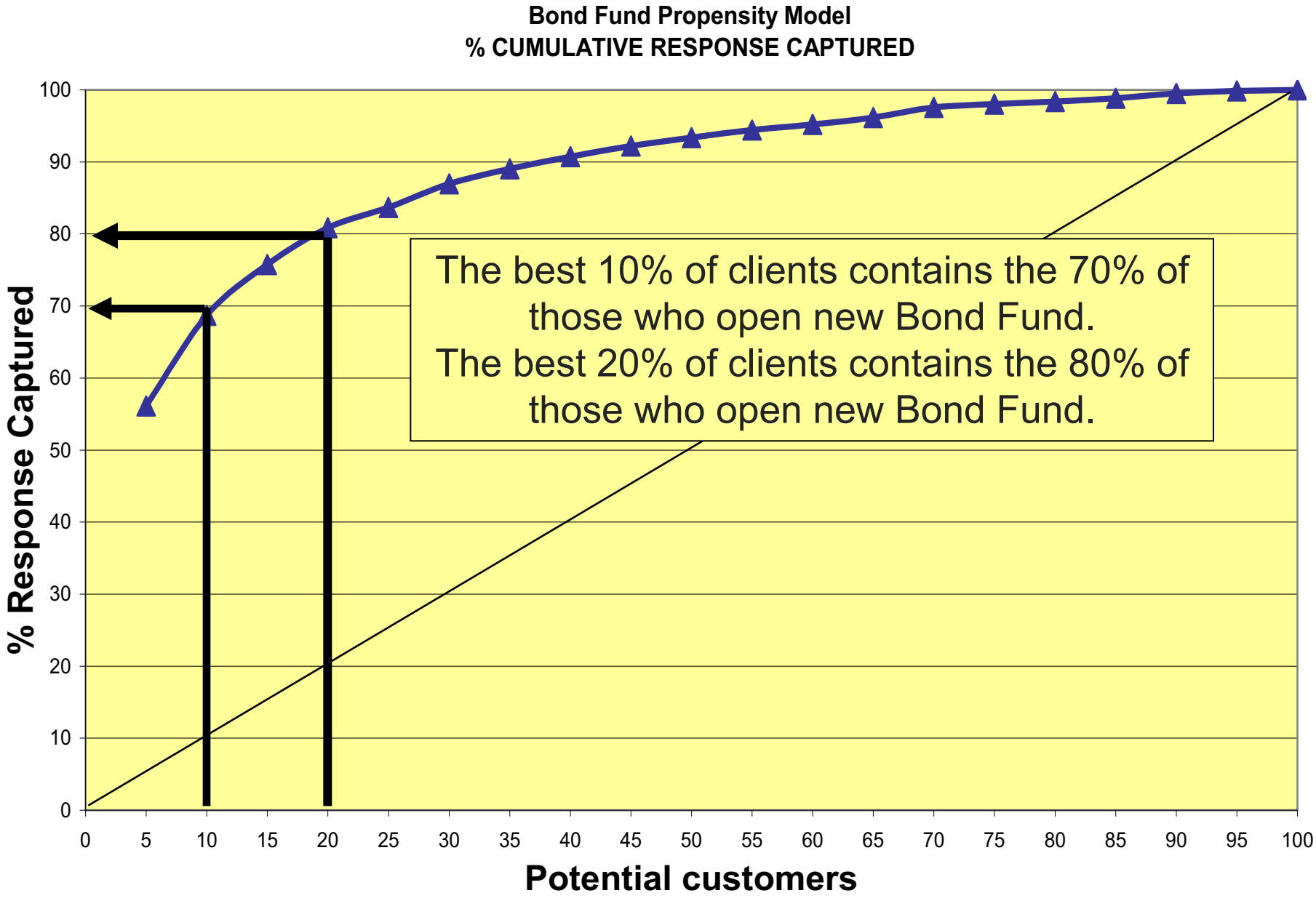
- A: variables that describe, for the product, details about the behavior of the client in date 30/09/2005
- B: variables that summarize the behavior of the client in date 30/09/2005 (all variables in input used for the segmentation have been considered)
- C: target variable, measures the opening of a Bond Fund between 01/11/2005 and 31/01/2006



Validation: curve lift and redemption wait



Validation: % of captured response



Campaign implementation(1/2)

% Population	Score MIN-MAX	Avg Score	Cum. # Customers	Cum. # Target	Cum. Redemption	Cum. Lift	% Response Captured
5	1:235-952	269	93,915	1,115	1.19	11.22	56.12
10	2:216-235	224	187,825	1,365	0.73	6.87	68.70
15	3:207-216	211	281,705	1,505	0.53	5.05	75.74
20	4:198-207	202	375,627	1,607	0.43	4.04	80.88
45	9:171-175	173	845,132	1,832	0.22	2.05	92.20
50	10:169-171	170	939,055	1,855	0.20	1.87	93.36
75	15:155-157	156	1,408,548	1,948	0.14	1.31	98.04
80	16:152-155	154	1,502,455	1,955	0.13	1.23	98.39
85	17:147-152	149	1,596,364	1,964	0.12	1.16	98.84
90	18:142-147	144	1,690,277	1,977	0.12	1.11	99.50
95	19:135-142	139	1,784,164	1,984	0.11	1.05	99.85
100	20:-108-135	116	1,878,087	1,987	0.11	1.00	100.00

If we consider the best 5% of clients (about 95,000 contacts), the redemption wait is 11.2 times the average redemption and we expect to catch the 56% of potential clients of new Bond Fund.

If we consider the best 10% of clients (about 190,000 contacts) the redemption wait is 6.9 times the average redemption and we expect to catch the 69% of potential clients of new Bond Fund.

Agenda

Module 5: Scoring Models

- Introduction
- Scoring Models & Marketing Campaigns
- The Analytical Process

Logistic Regression Model

The logistic regression is a type of Generalized Liner Model.

It allows to predict a discrete variable, that can be understood as the belonging to a group, on the basis of a set of variables (continuous, discrete, dichotomic).

Generally, the dependent variable (or response variable) is dichotomic and it represents absence/presence or failure/success.

Examples:

- Churn Model (event: abandonment)
- Propensity Model (event: purchase)

Logistic Regression Model

Model Hypothesis

Y	X_1	X_2	X_3	…	…	…	X_p
y_1	x_{11}	x_{12}	x_{13}	…	…	…	x_{1p}
y_2	x_{21}	x_{22}	x_{23}	…	…	…	x_{2p}
y_3	x_{31}	x_{32}	x_{33}	…	…	…	x_{3p}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
y_n	x_{n1}	x_{n2}	x_{n3}	…	…	…	x_{np}
(nx1)				(nxp)			

- n statistical units
- column vector (nx1) of n measurements on a dichotomic variable (Y)
- matrix (nxp) of n measurements on p quantitative variables ($X_1,\ldots,X_p$)
- the single observation is the row vector ($y_i, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}$) i=1,…,n

Logistic Regression Model

Model Hypothesis

The dipendent dichotomic variable Y denotes the presence or the absence of a specific feature.

Y takes value 1 with probability π and value 0 with probability 1-π.

Y is distributed as a **bernoullian** random variable with parameter π, that describe the outcome of a random experiment that can succeed with probability π.

$$Y \sim Bernoulli(\pi)$$

$$Pr(Y) = \pi^{Y}(1-\pi)^{(1-Y)}$$

Logistic Regression Model

Model Hypothesis

The Linear regression model is unsuitable when the response variable is dichotomic because:

1. It doesn't guarantee the observance of the range [0,1]
2. The error component can assume only two values, it can't have a normal distribution.
3. The error component violates the homoscedasticity hypothesis, the variance depends on the specific value of Xi

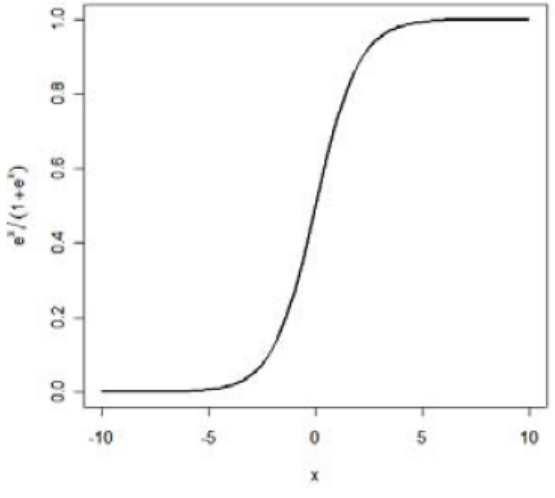
Logistic Regression Model

Model Hypothesis

To apply the logistic regression, it is assumed that π: Pr(Y=1 I X) is defined as follows:

$$\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i = \frac{\exp(\underline{X}_i^T \underline{\beta})}{1 + \exp(\underline{X}_i^T \underline{\beta})}$$

Logistic Function



Logistic Regression Model

Model Hypothesis

The logistic model has some important properties:

1. It respects the restriction that the expected value of: $\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i$ is included in the interval [0,1];
1. The «S» shape of the logistic function guarantees a gradual approach to the extreme values 0 and 1;
2. The logit function of: π_i can be expressed as a linear combination of the independent variables X_1,.., X_k:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

Logistic Model Regression

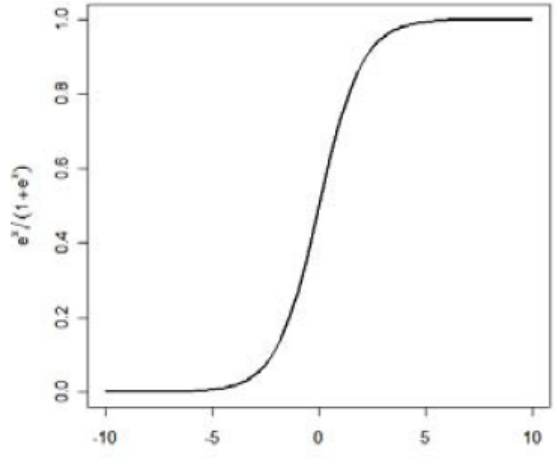
Model Hypothesis

Let:

$$\Pr(Y_i = 1 \mid \underline{X}_i) = \pi_i$$

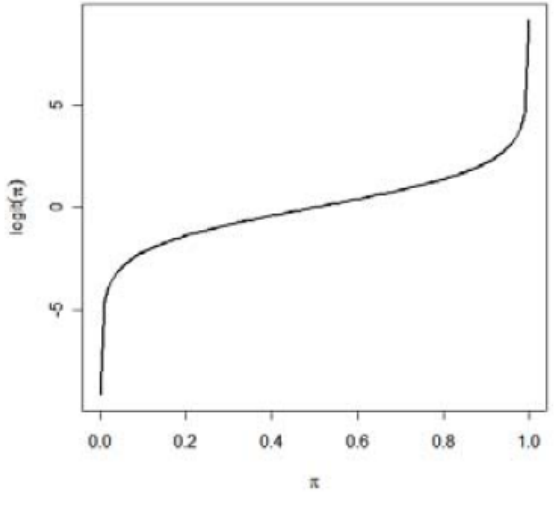
It can be shown that

$$\pi_i = \frac{\exp(\underline{X}_i^T \underline{\beta})}{1 + \exp(\underline{X}_i^T \underline{\beta})} \qquad \text{LOGISTIC}$$



is equivalent to

$$\log\left(\frac{\pi_i}{1 - \pi_{i\,i}}\right) = \underline{X}_i^T \underline{\beta} \qquad \text{LOGIT}$$



Logistic Regression Model

Model Estimate

Similarly to the linear regression model, the relationship between the dependent variable and the independent ones is known except for the values of the parameters:

$$\mathrm{logit}\ (\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

We need a method that allows to obtain a “good” estimation of the parameters on the basis of the available sampling observations.

Logistic Regression Model

Model Estimate

It can be shown that the estimators obtained through the Least Squares method don't satisfy the optimal properties that the linear regression owns.

The more general Maximum Likelihood Method shall be used. It is based on the maximization of the probability of observing the avalaible sampling dataset as a function of β.

- The likelihood equations are not linear in the parameters and they do not permit an explicit solution (except for specific cases).

- It is necessary to use iterative numeric methods to approximate the solution (Newton-Raphson's or Scoring's Fisher's Algorithm)

Logistic Regression Model

Model Estimate

The maximum likelihood estimators have optimal properties when the sampling size is numerically big:

- asymptotically unbiased (the estimates are unbiased, they approach the true value)
- asymptotically efficient (the standard error of the estimate is at least as small as the standard error of any other estimate model)
- asymptotically normal (it's possible to use the normal or chi-squared distribution to compute the confidence intervals)

Logistic Regression Model

Model Estimate

Test to evaluate the joint significance of the coefficients ("Testing Global Null Hypothesis: BETA=0")

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

- Likelihood Ratio
- Score
- Wald

These statistics have a chi-squared distribution with *n* degrees of freedom, where n is the number of estimated coefficients of the independent variables.

If the p-value is small (reject H_0), then the model has a good explanatory capacity.

N.B. Equivalent to F Test in the linear regression

Logistic Regression Model

Model Estimate

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2192.4978	7	<.0001
Score	1399.0552	7	<.0001
Wald	876.2357	7	<.0001

Logistic Regression Model

Model Estimate

Summary indicators of the goodeness of the model

- Likelihhod ratio test → OK p-value with small values
  → It's the analogous of the F test in the linear regression

- Wald Chi_square test → OK p-value with small values
  → It's the analogous of the t test in the linear regression

- Akaike Criterion → OK small values

- Schwart Criterion → OK small values

Logistic Regression Model

Model Evaluation

Let PAIRS be the number of pairs of observation (i,h with i≠h) that in a case have Y=1 and in the other case Y=0.

The pair of observations (i,h con i≠h) for which $Y_i = 1$ and $Y_h = 0$ is:

- concordant if $\hat{\pi}_i > \hat{\pi}_h$
- tied if $\hat{\pi}_i = \hat{\pi}_h$
- discordant if $\hat{\pi}_i < \hat{\pi}_h$

The bigger the number of CONCORDANT is (and then the smaller the number of DISCORDANT is), the better the model will describe the investigated event.

Logistic Regression Model

Model Evaluation

The following statistics are computed on the basis of the number of CONCORDANT, DISCORDANT and TIED pairs.

$$Tau-a = \frac{C-D}{N}$$

$$Gamma = \frac{C-D}{C+D}$$

$$Somer'sD = \frac{C-D}{C+D+T}$$

$$c = 0.5*(1+Somer'sD)$$

Association of Predicted Probabilities and Observed			
Percent Concordant	89.6	Somers' D	0.796
Percent Discordant	10.0	Gamma	0.800
Percent Tied	0.4	Tau-a	0.146
Pairs	643691936	c	0.898

Where:

- C is the number of CONCORDANT pairs,
- D is the number of DISCORDANT pairs,
- T is the number of TIED pairs,
- N is the number of pairs.

The bigger these indicators are, the more the model is “correct”. These measurements vary between 0 and 1. Greater values correspond to a stronger connection between predicted values and observed values.

Logistic Regression Model

Model Estimate

Test to evaluate the significance of the single coefficients:

$$H_0 : \beta_j = 0$$

– Wald Chi-square: the square of the ratio of the estimate and the standard error

The coefficient is significally different from zero if the corresponding p-value is small (that is, we reject the hypothesis of null coefficient) → the regressor to which the coefficient is associated is relevant to the event explanation.

N.B. Equivalent to the t Test in the linear regression

Logistic Regression Model

Model Estimate

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.2530	0.1147	119.3602	<.0001	
PAG_ORD	1	0.000070	5.295E-6	175.1845	<.0001	1.1035
TOT_ORD	1	0.5151	0.0432	142.1610	<.0001	0.6494
PAG_MES	1	0.000120	8.608E-6	194.9225	<.0001	0.6074
SUD	1	-0.8965	0.1038	74.6650	<.0001	-0.2381
CEN	1	-0.2745	0.1294	4.5039	0.0338	-0.0571
SESSO	1	0.2729	0.1005	7.3780	0.0066	0.0695
LISTA	1	-0.00293	0.0553	0.0028	0.9577	-0.00134

Logistic Regression Model

Model Estimate

When we have quantitative regressors, the standardized coefficients can be useful to evaluate the relative importance of the variables and to understand which one influence the model more.

Relatively to the last example:

- The most influencial variable in the model is PAG_ORD (Standardized estimated: 1.1035),
- follows TOT_ORD (Standardized estimated: 0.6494),
- follows PAG_MES (Standardized estimated: 0.6074), etc.

Logistic Regression Model

Model Estimate

Similarly to the linear regression model, the relationship between the dependent variable and the independent ones is known except for the values of the parameters:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

For the purposes of the formulation of a linear model it is necessary to:

1. turn the probability in odds π/(1- π) in order to remove the upper limit (Sup=1)
2. apply the logarithmic function to the odds in order to remove the lower limit (Inf=0)

Logistic Regression Model

Model Interpretation

In betting it is said that a certain event is given 5 to 2, that means that the odds is 5/2: the ratio between the expected number of times in which the event happens and the expected number in which the event doesn't happen.

There's a simple relationship between odds and probability:

$$O = \frac{\pi}{1-\pi}$$

$$\pi = \frac{O}{1+O}$$

where π is the event probability and O is the odds.

Logistic Regression Model

Model Interpretation

An odds below 1 corresponds to a probability below 0.5. The lower limit is 0 as for the probability, but the odds doesn't have upper limit.

Event probability	odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Logistic Regression Model

Model Interpretation

In the logistic regression, a coefficient 0.2 means that the logit of Y (the log of the odds) increase of 0.2 at the hold of feature X. What does the 0.2 increase of logit mean?

Since the relationship between probability and regressor is not linear, it's easier to talk in odds terms. The estimated coefficients, apart from the sign, are not interpretable, but the odds ratio (the exp. of the coeff.) does.

Example (Churn Model):

Sesso	Estimate	CHURN RATE	Odds Ratio Estimate
0 (femmina)		1.98%	
1(maschio)	0.2103	2.52%	1.23
TOTAL		2.24%	

Males have a churn rate greater than females.

The expected odds for male churn is 1.234 times the female one (23% higher).

Logistic Regression Model

Model Interpretation

Independent variable (ex. M=1; F=0)

Response variable (YES=1; NO=0)	x = 1	x = 0
y = 1	π(1)	π(0)
y = 0	1- π(1)	1- π(0)

ODDS RATIO

$$\psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{ODDS_1}{ODDS_0}$$

It's an association measure; it approximates the Relative Risk, that is how much is probable for the response variable to have x=1 respect to have x different from 0.

Logistic Regression Model

Model Interpretation

In case of continuous variables, the parameter interpretation is analogous.

The coefficients expresses the logit improvement at unitary increase of X.

$$\beta_1 = logit(Pr(Y = 1 | X = x + 1) - logit(Pr(Y = 1 | X = x)$$

Logistic Regression Model

Model Interpretation

Odds Ratio Estimates

Effect	Point Estimate
PAG_ORD	1.000
TOT_ORD	1.674
PAG_MES	1.000
SUD	0.408
CEN	0.760
SESSO	1.314
LISTA	0.997

Logistic Regression Model

Model Evaluation

Similarly to linear regression, also for the logistic regression the problem of multicollinearity can cause side effects on estimation stability.

The operating methods of the problems are similar to the ones of the linear regression.

Logistic Regression Model

Model Evaluation

Similarly to linear regression, it's possible to use different methods of automatic selection of the variables.

Also in this case, the algorithms works according to the logic of:

- Stepwise
- Forward
- Backward

Logistic Regression Model

Model Application

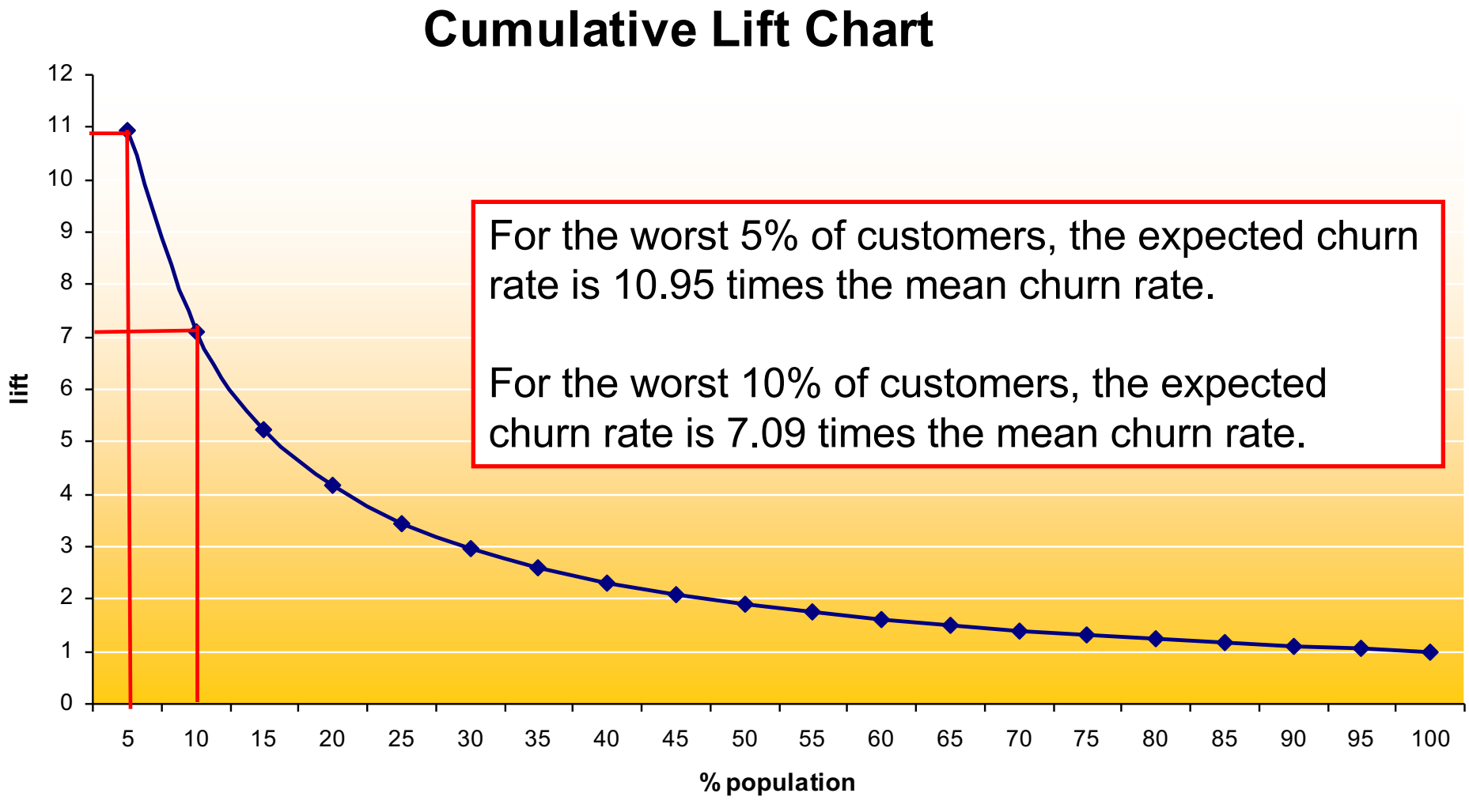
All the observations are divided in ventiles on the basis of expected response probability.

decili	target	popolazione	target cumulato	popolazione cumulata	redemption	redemption cumulata	lift	lift cumulata	%catturati	%catturati cumulata
5	1028	4191	1028	4191	24.53%	24.53%	10.95	10.95	54.76%	54.76%
10	303	4191	1331	8382	7.22%	15.88%	3.23	7.09	16.13%	70.88%
15	144	4191	1475	12573	3.44%	11.73%	1.54	5.24	7.68%	78.57%
20	85	4191	1560	16764	2.02%	9.30%	0.90	4.15	4.51%	83.08%
25	62	4191	1622	20955	1.48%	7.74%	0.66	3.46	3.31%	86.39%
30	50	4191	1672	25146	1.18%	6.65%	0.53	2.97	2.64%	89.03%
35	35	4191	1707	29337	0.84%	5.82%	0.38	2.60	1.88%	90.91%
40	29	4191	1736	33528	0.69%	5.18%	0.31	2.31	1.54%	92.46%
45	25	4191	1761	37719	0.60%	4.67%	0.27	2.08	1.33%	93.79%
50	23	4191	1784	41910	0.55%	4.26%	0.24	1.90	1.22%	95.01%
55	17	4191	1801	46101	0.41%	3.91%	0.18	1.74	0.92%	95.93%
60	16	4191	1817	50292	0.37%	3.61%	0.17	1.61	0.83%	96.76%
65	13	4191	1830	54483	0.31%	3.36%	0.14	1.50	0.69%	97.46%
70	11	4191	1840	58674	0.25%	3.14%	0.11	1.40	0.57%	98.02%
75	6	4191	1847	62865	0.15%	2.94%	0.07	1.31	0.33%	98.35%
80	6	4191	1853	67056	0.15%	2.76%	0.07	1.23	0.33%	98.68%
85	6	4191	1859	71247	0.15%	2.61%	0.07	1.16	0.33%	99.01%
90	6	4191	1865	75438	0.15%	2.47%	0.07	1.10	0.33%	99.34%
95	6	4191	1871	79629	0.15%	2.35%	0.07	1.05	0.33%	99.67%
100	6	4191	1878	83820	0.15%	2.24%	0.07	1.00	0.33%	100.00%

The Lift value is obtained as the ratio between the percentage of positives contained in the ventiles and the percentage of positives contained in the total population.

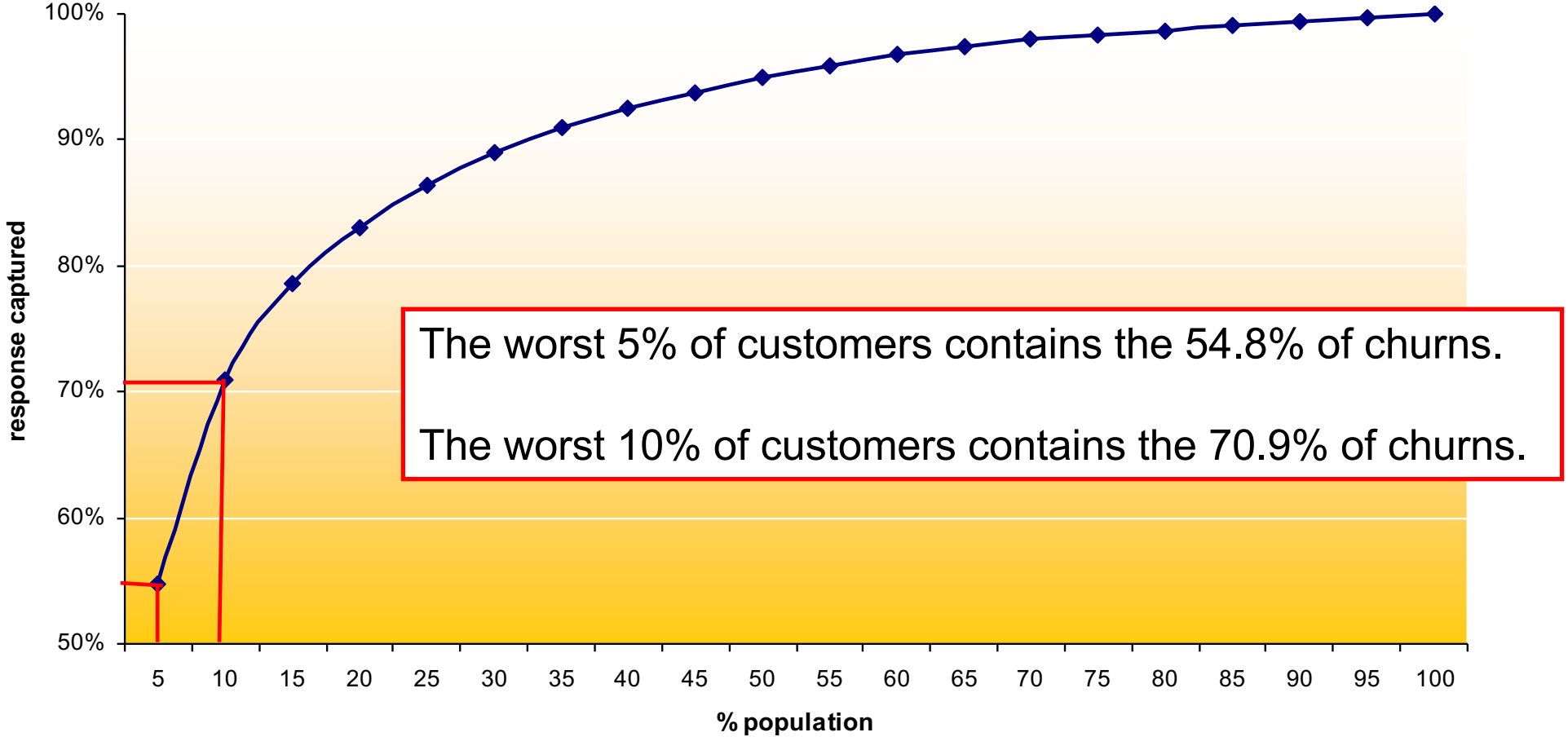
Logistic Regression Model

Model Application



Logistic Regression Model

Model Application



Key points(1/3)

Critical success factors in Customer Profiling projects are:

- ability to design a complete input data source, considering the availability of data and a scope of dimensions large enough to allow us to measure the clients' behaviour.

- a solid methodology, based on projects experience, aimed to produce concrete results and useful to support the business.

Key points(2/3)

In CP projects, the best input data for analysis are data actually available and not those we would like to have available:

- that means we must have an extremely concrete and “business driven” approach.

Key points(3/3)

The efficiency of the output produced depends on the quality and the wealth of information, as well as a correct translation of the business goals in analysis goals:

- do not ignore elementary analysis to pursue the construction of complex models.