Measure Responses of Customers

(1) x1:Gender, (2) x2:Hotline indicator 1, (3) x3:Hotline Indicator 2
The firm has sent out solicitations to 200 persons.

	x1	x2	x3	y		x1	x2	x3	y		x1	x2	x3	y	x1	x2	x3	y	
1	0	99	73	1	11	1	16	71	0	21	1	88	23	1	31	0	3	63	0
2	1	13	82	0	12	1	23	14	0	22	0	40	62	0	32	1	82	96	1
3	0	49	58	0	13	0	93	66	0	23	1	57	14	0	33	0	50	27	0
4	1	2	97	0	14	0	75	31	0	24	0	6	90	0	34	1	97	6	0
5	1	93	39	0	15	0	48	11	0	25	1	85	73	1	35	1	29	19	0
6	0	26	8	0	16	1	65	50	0	26	1	87	98	1	36	1	19	69	0
7	0	7	62	0	17	0	41	76	0	27	1	40	53	0	37	0	89	20	0
8	1	10	10	0	18	0	61	13	0	28	1	69	54	0	38	1	77	25	0
9	0	91	47	0	19	1	28	82	0	29	0	35	48	0	39	0	31	72	0
10	1	72	84	0	20	0	16	59	0	30	0	62	72	0	40	0	38	4	0

Characteristic

Start with summary statistics:

Sample size = 200

	x1	x2	x3	y
Mean	0.52	50.095	51.835	0.15
Std. Dev	0.501	29.263	29.177	0.358

Goal: To forecast y as a function of the variables we have data on: x1, x2, x3

Logistic Regression

- Logistic regression: y only takes value 0 and 1

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$p(y = 1 \mid t) = \frac{\exp(t)}{1+\exp(t)}$$

- Run logistic regression on the training data

Coefficients:

	Value	Std. Error
(Intercept)	-12.39913750	2.12044072
x1	1.33124053	0.64350879
x2	0.08114589	0.01625986
x3	0.07261875	0.01557379

Prediction

- Suppose the firm has 300 further addresses that you are considering for customer acquisition. Who should the firm contact?

	x1	x2	x3		x1	x2	x3		x1	x2	x3		x1	x2	x3
1	1	64	11	11	1	66	73	21	0	21	45	31	0	9	58
2	1	30	32	12	1	15	95	22	1	67	71	32	1	29	65
3	0	49	60	13	1	42	81	23	1	2	70	33	1	14	45
4	1	23	42	14	0	46	25	24	0	12	15	34	1	91	42
5	1	94	63	15	0	9	15	25	1	26	37	35	1	99	6
6	1	52	28	16	1	33	99	26	1	13	79	36	1	52	54
7	0	38	75	17	1	5	74	27	1	28	62	37	1	30	96
8	1	43	23	18	0	49	24	28	0	33	53	38	1	12	50
9	0	49	1	19	0	18	68	29	1	40	58	39	1	84	22
10	1	74	68	20	1	68	16	30	1	44	40	40	1	39	1

List Scoring

- Compute the following for each person in the list
  - The value of *t*

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

  - Predicted response probability

$$p(y=1 \mid t) = \frac{\exp(t)}{1+\exp(t)}$$

  - Predicted lift

$$\text{Lift} = \frac{p(y=1 \mid x_1, x_2, x_3)}{p(\text{response} \mid \text{population})} = \frac{p(y=1 \mid x_1, x_2, x_3)}{.15} \text{ in this case}$$

  - The firm ought to more interested in persons with the higher values of above measures

List Scoring

- The three measures turn out to be the following:

	x1	x2	x3	score	py	lift
1	1	64	11	-5.076	0.006	0.041
2	1	30	32	-6.310	0.002	0.012
3	0	49	60	-4.066	0.017	0.112
4	1	23	42	-6.152	0.002	0.014
5	1	94	63	1.135	0.757	5.045
6	1	52	28	-4.815	0.008	0.054
7	0	38	75	-3.869	0.020	0.136
8	1	43	23	-5.908	0.003	0.018
9	0	49	1	-8.350	0.000	0.002
10	1	74	68	-0.125	0.469	3.125
11	1	66	73	-0.411	0.399	2.658
12	1	15	95	-2.952	0.050	0.331

- Over all 300 persons, average “py” is 0.126, average “lift” is 0.8408

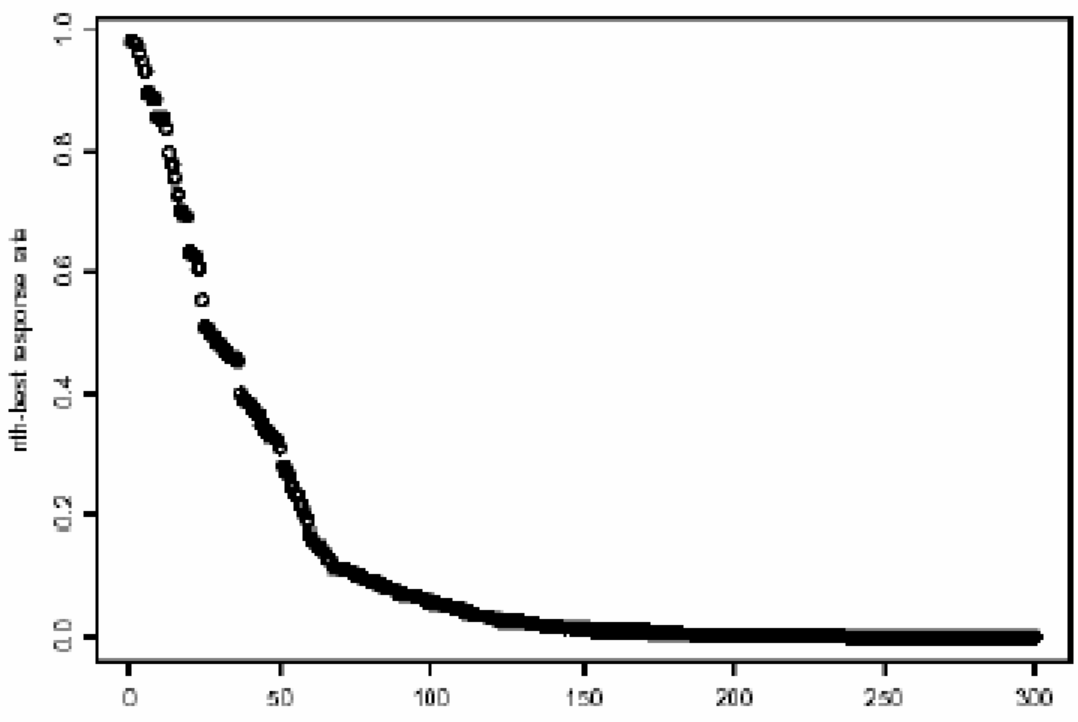
List Scoring

- Sort the prospects in descending order

	x1	x2	x3	score	py	lift		x1	x2	x3	score	py	lift
223	1	97	99	3.99	0.982	6.55	48	0	1	38	-9.56	7.06E-05	0.000471
238	1	100	93	3.8	0.978	6.52	287	0	23	13	-9.59	6.85E-05	0.000457
104	1	97	89	3.27	0.963	6.42	90	1	11	7	-9.67	6.33E-05	0.000422
192	1	83	100	2.93	0.949	6.33	125	0	3	31	-9.90	4.99E-05	0.000333
161	1	82	97	2.63	0.933	6.22	258	0	19	13	-9.91	4.95E-05	0.00033
184	1	94	77	2.15	0.896	5.97	55	0	15	16	-10.00	4.45E-05	0.000297
185	1	98	72	2.11	0.892	5.95	296	0	21	9	-10.00	4.36E-05	0.00029
123	0	93	95	2.05	0.886	5.9	138	1	3	9	-10.20	3.83E-05	0.000255
143	0	87	98	1.78	0.855	5.7	24	0	12	15	-10.30	3.24E-05	0.000216
78	1	75	93	1.77	0.855	5.7	15	0	9	15	-10.60	2.54E-05	0.00017
271	1	74	94	1.76	0.854	5.69	208	0	14	2	-11.10	1.48E-05	0.000099
96	0	96	86	1.64	0.837	5.58	191	0	2	9	-11.60	9.32E-06	0.0000621

Curve for Marginal Response Rate vs Number of Solicitations Made

- Consider n solicitations are made to the n "best" prospects. Plot each value of n against the n^{th} highest predicted response rate



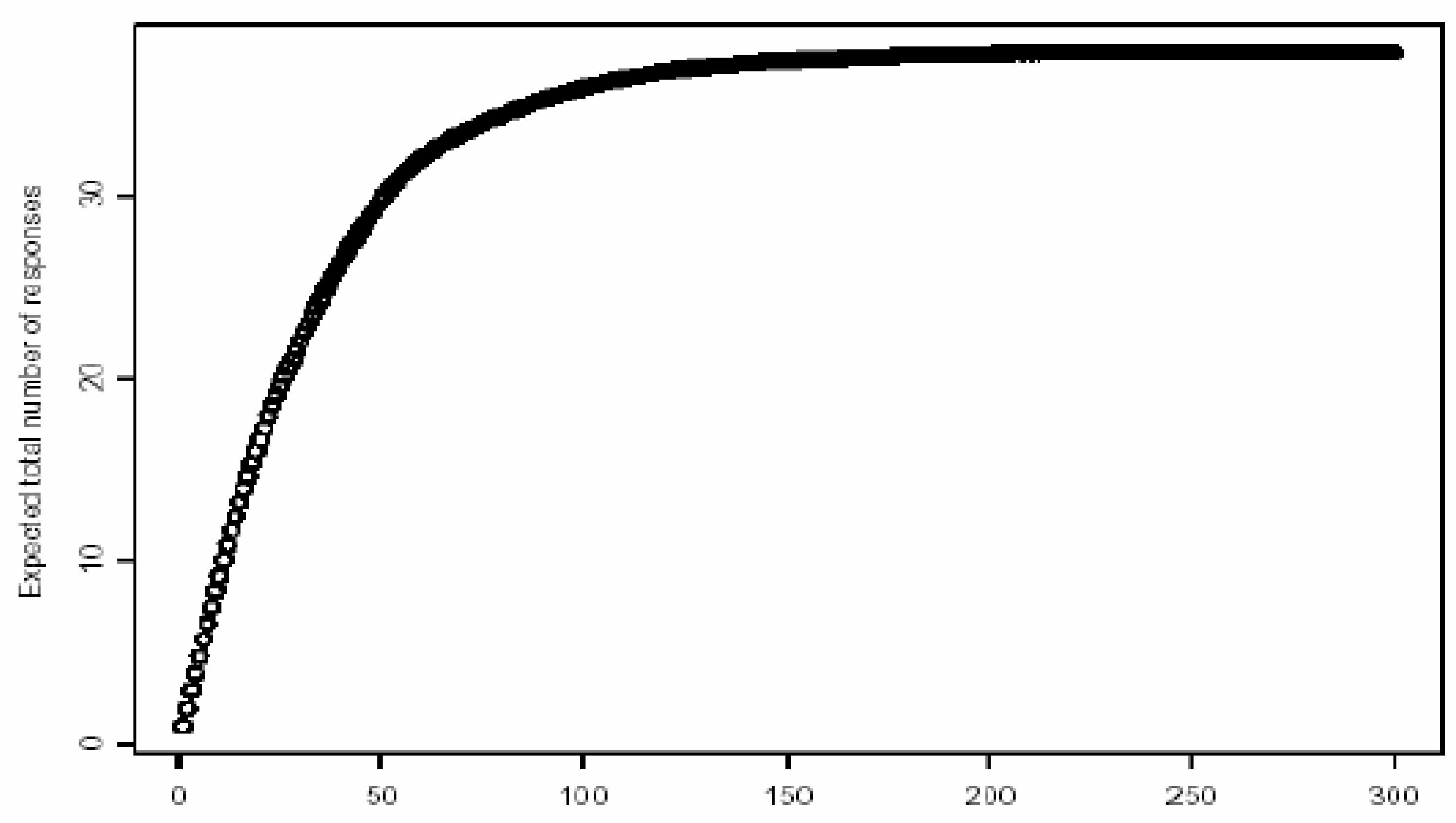
Curve for Number of Positive Responses vs Number of Solicitations Made

- If n solicitations are made to the n "best" prospects, the expected number of positive responses (sales) is the cumulative sum of the n highest values of p(y=1)

n	p(y=1)	Cumsum
1	0.982	0.982
2	0.978	1.96
3	0.963	2.92
4	0.949	3.87
5	0.933	4.81
6	0.896	5.7
7	0.892	6.59
8	0.886	7.48
9	0.855	8.33
10	0.855	9.19
11	0.854	10
12	0.837	10.9
...		
300 2	9.32E-06	37.801

Note that 300 x 0.126 = 37.8

Curve for Number of Positive Responses vs Number of Solicitations Made



Rollout run on best prospects

- Whom should the rollout run be deployed on? How many in rollout?
  - Limited supply rule: Suppose firm has only k items. Then firm should choose n such that the sum of n highest predicted response rate (cumsum()) is just under k.
  - Example, for k=20, cumsum(25)=19.57 and cumsum(26)=20.07. Therefore, the firm should mail to the 25 highest prospects.

Model Assessment via Holdout Data

- Assume that the firm had actual response data for the second list of 300 persons.
- Examine the “confusion” matrix, which is the cross-tab of predicted response against actual response.
- Note: Predicted response is 0 if response rate is less than 0.5 and 1 otherwise

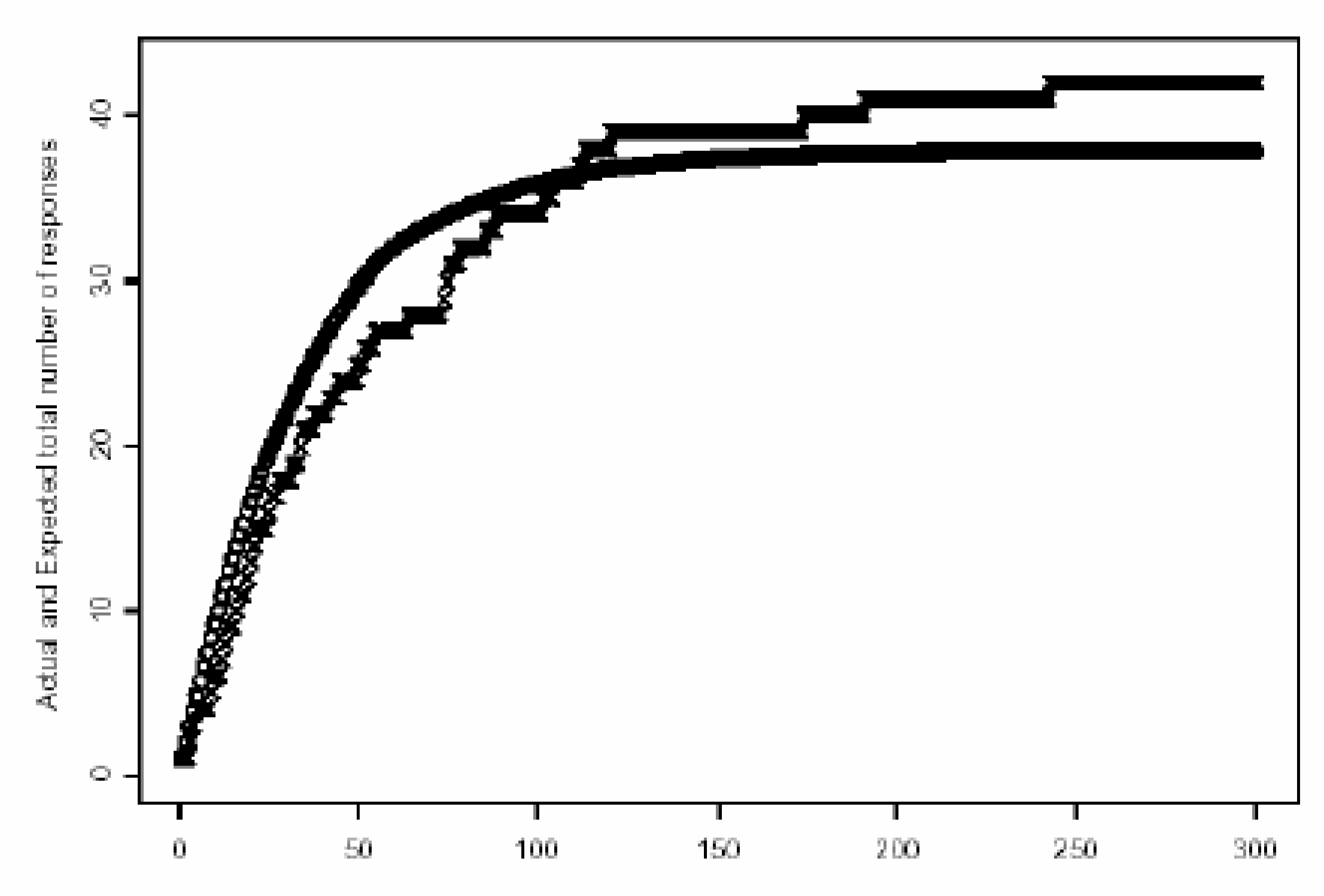
		Predicted y	
		0	1
Actual y	0	249	9
	1	25	17

Total error rate is $\frac{9+25}{300}=.1133$

Holdout Data Analysis

n	x1	x2	x3	score	py	pred Y	act.Y	cum(act.Y)	cum(act.Y)/n
1	1	97	99	3.992	0.982	1	1	1	1.000
2	1	100	93	3.800	0.978	1	0	1	0.500
3	1	97	89	3.266	0.963	1	1	2	0.667
4	1	83	100	2.929	0.949	1	1	3	0.750
5	1	82	97	2.630	0.933	1	1	4	0.800
6	1	94	77	2.151	0.896	1	0	4	0.667
7	1	98	72	2.113	0.892	1	0	4	0.571
8	0	93	95	2.046	0.886	1	0	4	0.500
9	0	87	98	1.777	0.855	1	1	5	0.556
10	1	75	93	1.771	0.855	1	1	6	0.600
11	1	74	94	1.763	0.854	1	0	6	0.545
12	0	96	86	1.636	0.837	1	1	7	0.583

Holdout Data Analysis



Example

- Marketing Problem: catalog company wants to acquire new customers
- Method: use a rented list, e.g., from another catalog company
- Predict: will a customer on this list respond if we send an offer?
- One approach
  - Perform a test mailing on the list
  - Build data-mining model linking response to other information in the database
  - Apply model to entire database

Linear Regression Model

- Linear regression model:

$$y = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3 + \varepsilon$$

$$p(y = 1 \mid t) = B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3$$

	Value	Std. Error
(Intercept)	-0.42300	0.060700
x1	0.04620	0.040900
x2	0.00579	0.000703
x3	0.00500	0.000706

py.linprob = -0.42300 + 0.04620*x1 + 0.00579*x2 + 0.00500*x3

List Scoring with the Linear Regression Model

	x1	x2	x3	py.linprob	lift.linprob
1	1	64	11	0.0483	0.322
2	1	30	32	-0.0434	-0.289
3	0	49	60	0.16	1.07
4	1	23	42	-0.0339	-0.226
5	1	94	63	0.482	3.21
6	1	52	28	0.0639	0.426
7	0	38	75	0.172	1.15
8	1	43	23	-0.0132	-0.0879
9	0	49	1	-0.135	-0.899
10	1	74	68	0.391	2.61
11	1	66	73	0.37	2.47
12	1	15	95	0.185	1.23

List Scoring with the Linear Regression Model

Sort the prospects in decreasing order of lift:

	x1	x2	x3	py.linprob	lift.linprob		x1	x2	x3	py.linprob	lift.linprob
223	1	97	99	0.68	4.53	125	0	3	31	-0.251	-1.67
238	1	100	93	0.667	4.45	55	0	15	16	-0.257	-1.71
104	1	97	89	0.63	4.2	296	0	21	9	-0.257	-1.71
192	1	83	##	0.604	4.02	152	1	12	9	-0.263	-1.75
123	0	93	95	0.59	3.94	205	1	1	21	-0.266	-1.78
161	1	82	97	0.583	3.89	289	1	13	7	-0.267	-1.78
143	0	87	98	0.571	3.8	90	1	11	7	-0.278	-1.86
96	0	96	86	0.563	3.75	24	0	12	15	-0.279	-1.86
184	1	94	77	0.552	3.68	15	0	9	15	-0.296	-1.98
185	1	98	72	0.55	3.67	138	1	3	9	-0.315	-2.1
134	0	85	93	0.534	3.56	208	0	14	2	-0.332	-2.22
78	1	75	93	0.522	3.48	191	0	2	9	-0.367	-2.45

Curve for Marginal Response Rate vs Number of Solicitations Made

- Consider n solicitations are made to the n "best" prospects. Plot each value of n against the n^{th} highest predicted response rate