Logistic Regression vs. Classification Trees

J D Freeman

May 21, 2014

Philanthropic Analytics Application
Using SAS JMP Pro

INDIANA UNIVERSITY

Outline

- Modeling
  1. Selecting Classification Tool and Parameters
     - Logistic Regression
     - Classification Trees
  2. Classification Under Asymmetric Response and Cost
  3. Calculate Net Profit
  4. Draw Lift Curves (side-by-side)
  5. Best Model
- Testing

MODELING

Logistic Regression & Classification Trees

- I have produced a number of models with substantially similar predictive success:

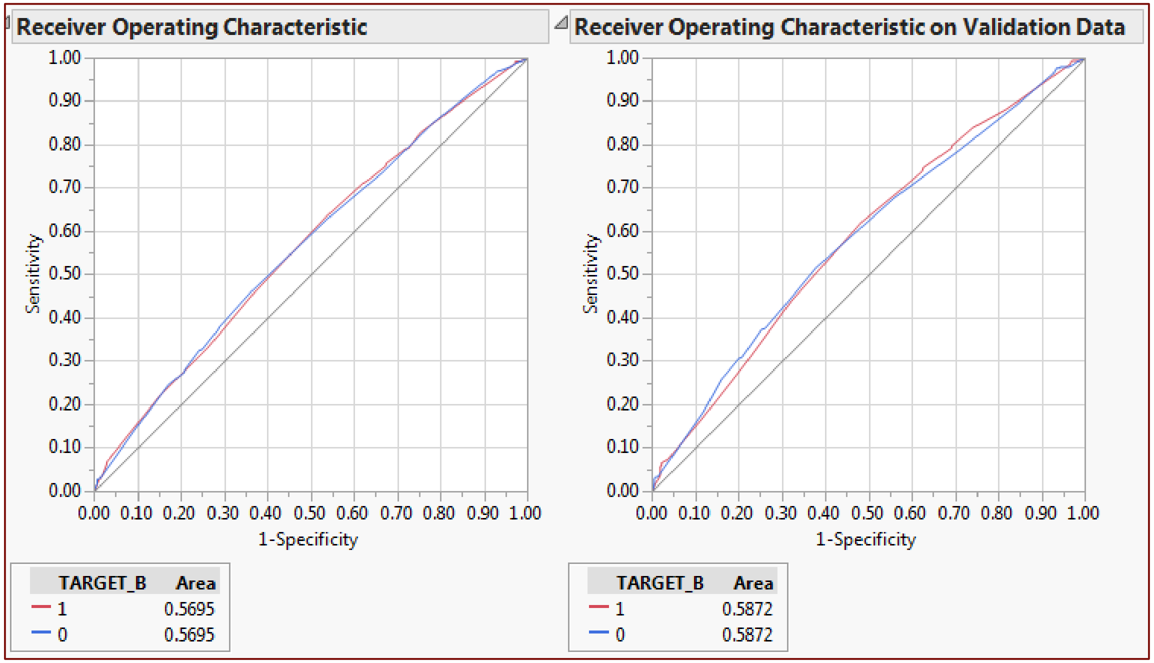
	Predictor Variables	Misclassification Rate Training	Misclassification Rate Validation	Misclassification Rate Weighted Misclassification	ROC AUC Training	ROC AUC Validation	Weighted ROC AUC	# Splits
Logistic Regression	genderdummy, SQRT_AVGGIFT*genderdummy	0.4567	0.4671	0.4609	0.5542	0.5376	0.5476	
	total months, NUMCHLD	0.4519	0.4455	0.4493	0.5695	0.5872	0.5766	
	total months	0.4546	0.4415	0.4494	0.5649	0.5876	0.5740	
Classification Trees	genderdummy, SQRT_AVGGIFT*genderdummy	0.4471	0.4696	0.4561	0.5585	0.5508	0.5554	5
	total months, NUMCHLD	0.4583	0.4952	0.4731	0.5601	0.5842	0.5697	3
	total months	0.4583	0.4255	0.4452	0.5601	0.5842	0.5697	3
	All Givens	0.4364	0.4679	0.4490	0.5575	0.5441	0.5521	2

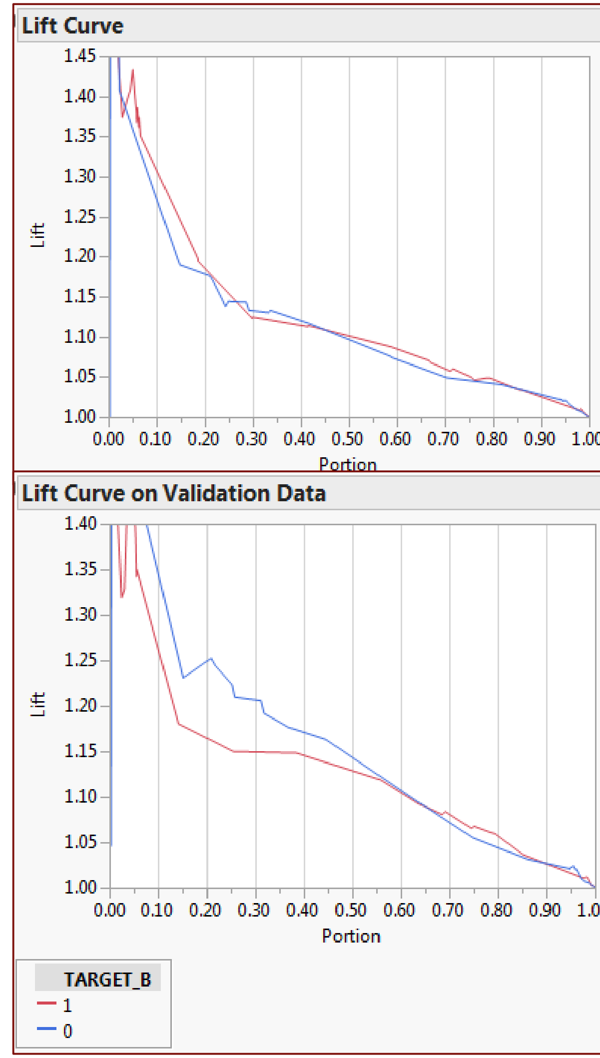
- I seek to minimize misclassification and maximize the area under the ROC curve.

- So there is a competition here between using Logistic Regression on (*totalmonths, NUMCHLD*) or a tree partition on only *totalmonths*.

Jump to Appendix

Logistic Regression – Sample JMP Results



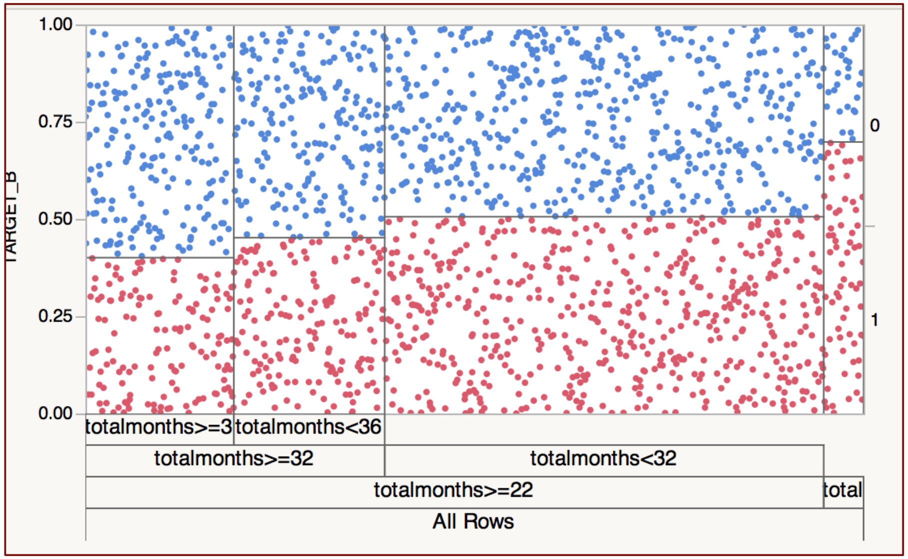


Confusion Matrix

Actual		Predicted
Training	**1**	**0**
1	418	490
0	356	608

Actual		Predicted
Validation	**1**	**0**
1	287	365
0	191	405

Classification Tree



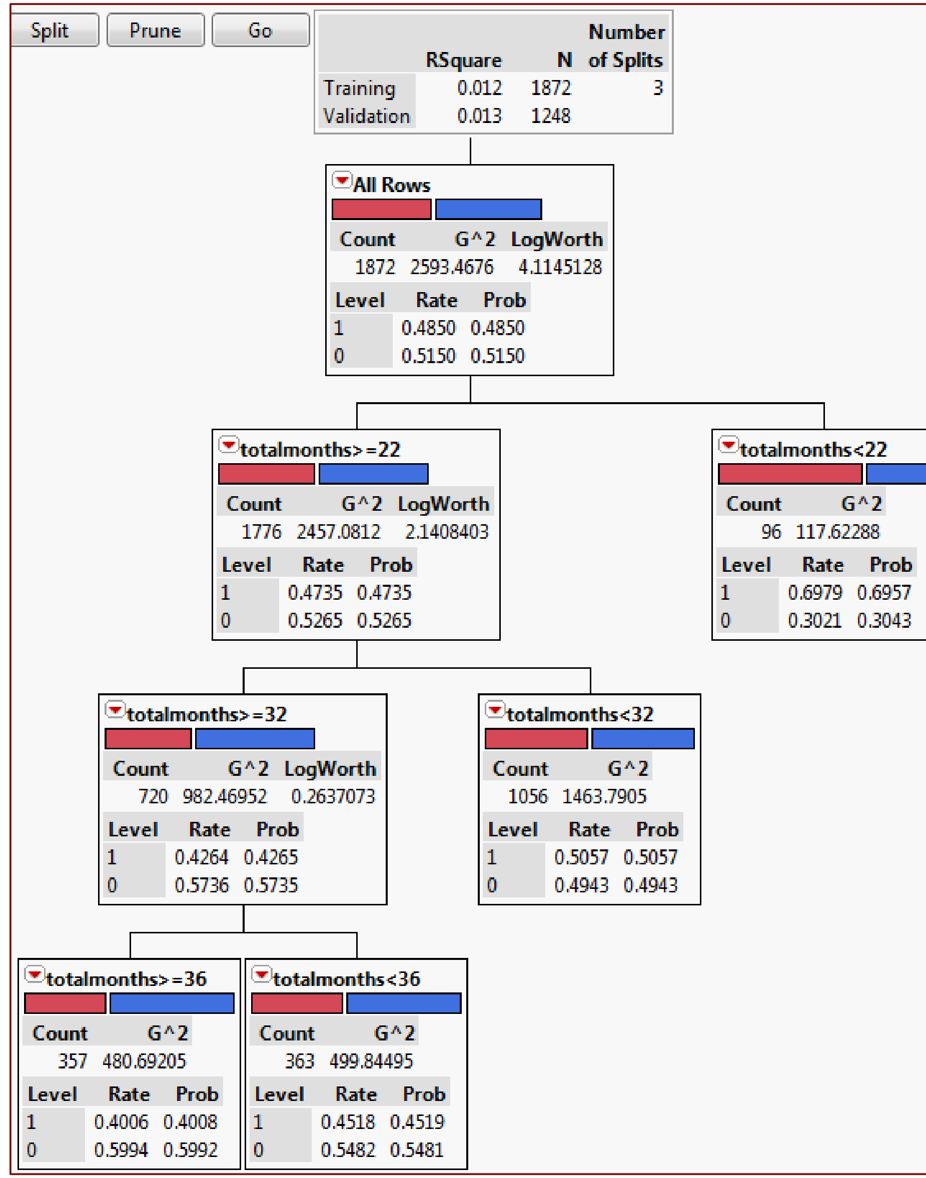
Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.0122	0.0126	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0223	0.0230	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6843	0.6834	∑ -Log(ρ[j])/n
RMSE	0.4956	0.4952	√∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4913	0.4914	∑ ly[j]-ρ[j]l/n
Misclassification Rate	0.4583	0.4255	∑ (ρ[j]≠ρMax)/n
N	1872	1248	n

Confusion Matrix

Actual Training	Predicted 1	Predicted 0
1	601	307
0	551	413

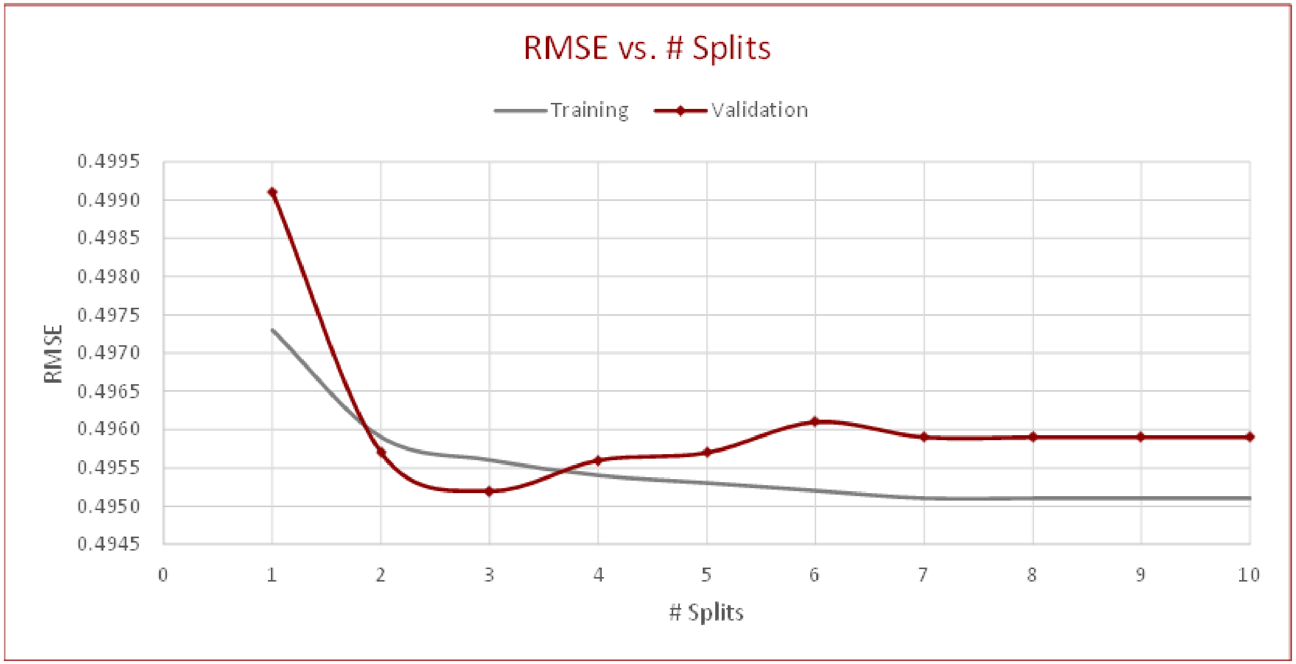
Actual Validation	Predicted 1	Predicted 0
1	427	225
0	306	290



Check for Overfitting

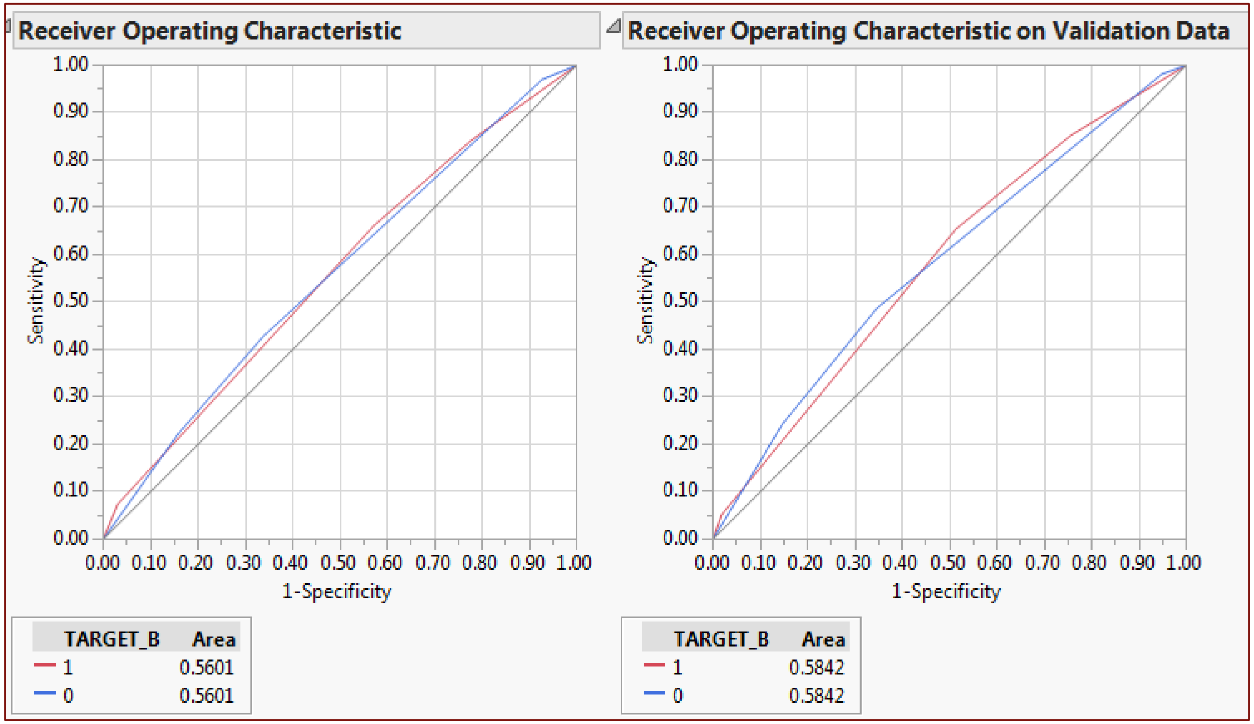
- Note the change in RMSE in the validation sets with the number of splits.

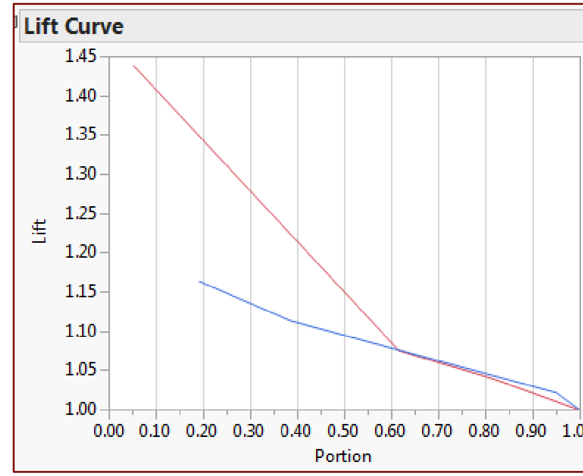
	RMSE	
# Splits	Training	Validation
1	0.4973	0.4991
2	0.4959	0.4957
3	0.4956	0.4952
4	0.4954	0.4956
5	0.4953	0.4957
6	0.4952	0.4961
7	0.4951	0.4959
8	0.4951	0.4959
9	0.4951	0.4959
10	0.4951	0.4959

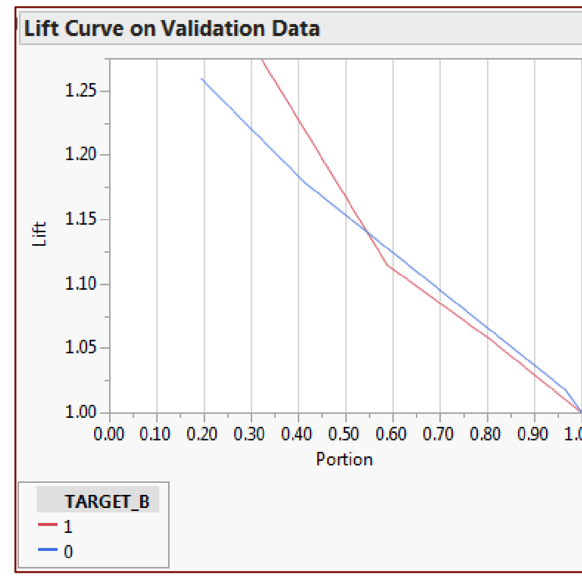


- After 3 splits, error rises in the validation set; further splits only benefit the training set.

Classification Tree – Sample JMP Results







Confusion Matrix

Actual / Predicted

Training	1	0
1	601	307
0	551	413

Actual / Predicted

Validation	1	0
1	427	225
0	306	290

Classification Under Asymmetric Response

- Use of a balanced sample (50% donors, 50% non-donors) is for the purpose of screening effects against randomness.
- If one produces random guesses on binary classification (e.g. flips a fairly weighted coin) enough times to get a statistical sampling of observations, the result would be evenly distributed between heads and tails.
- Any lift beyond 50% would then be the incremental predictive value of the model.
- Lift is the degree to which response is enhanced in a segment or class relative to the average population.

Classification Accuracy and Profit

- Classification accuracy is not necessarily the best performance metric for predicting profit.
- Our misclassification rates could be better, no matter which method is used.
- Profit is not only a function of the number of donors, but of how much each donor gives.
  - However, our costs are the same for each donor.
  - To maximize profit, maximize dollars collected, and minimize the number of donors solicited.
  - Examine cost/benefit of the included gifts.

Reweighting the Confusion Matrices

Logistic Regression

Confusion Matrix

Actual Training	Predicted 1	Predicted 0
1	418	490
0	356	608

Actual Validation	Predicted 1	Predicted 0
1	287	365
0	191	405

Tree Partition

Confusion Matrix

Actual Training	Predicted 1	Predicted 0
1	601	307
0	551	413

Actual Validation	Predicted 1	Predicted 0
1	427	225
0	306	290

The sample was balance at 50/50, but I expect only 5.1% donations in reality. The confusion matrices must be reweighted such that "1"s constitute only 5.1% of the total.

Training: 908+0.949X=X; X=17803.9216

Validation: 652+0.949X=X; X=12.784.3137

Logistic Regression (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	418	490	908
Actual 0	6,240	10,656	16,896
Total	6,658	11,146	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	287	365	652
Actual 0	3,889	8,244	12,133
Total	4,176	8,609	12,785

Tree Partition (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	601	307	908
Actual 0	9,658	7,238	16,896
Total	10,259	7,545	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	427	225	652
Actual 0	6,230	5,903	12,133
Total	6,657	6,128	12,785

Calculating Lift of Net Profit

For Logistic Regression (Training): Lift = 399.52%

Presume nothing is sent to predicted non-donors, saving cost:

Profit Before		# solicited	13,000,000
inflow	$8,619,000.00	response rate	5.10%
outflow	$8,840,000.00	average donation	$13.00
net	($221,000.00)	cost/solicitation	$0.68
Profit After (Regression - Training)		database size	13,000,000
inflow	$3,967,760.05	reweighted sample size	17804
outflow	$3,305,814.42	# solicited	6,658
net	$661,945.63	actual response	418
		solicitation rate	37.40%
LIFT:		actual response rate	6.28%
399.52%		average donation	$13.00
		cost/solicitation	$0.68

Calculating Lift of Net Profit

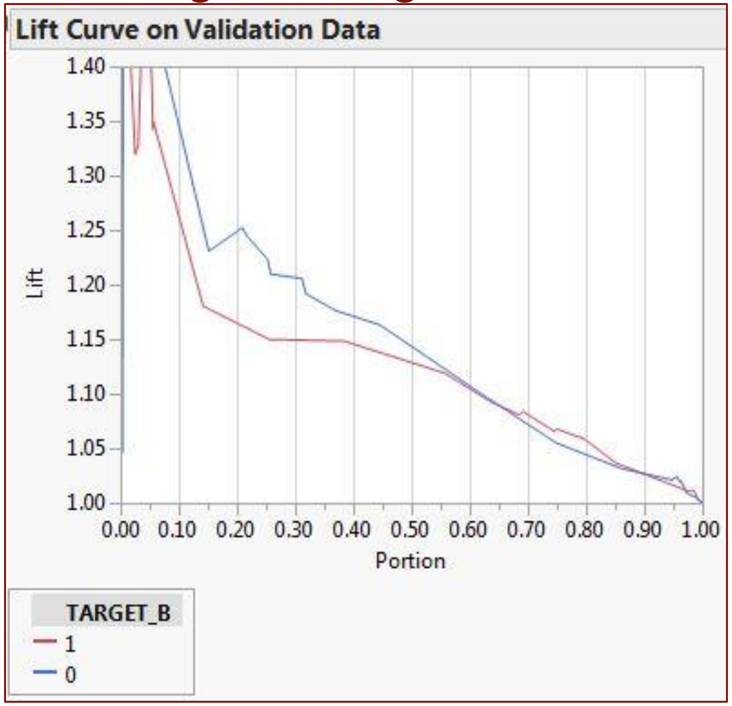
Similarly:

Method and Set	Lift
Logistic Regression, Training	399.52%
Logistic Regression, Validation	510.09%
Tree Partition, Training	376.50%
Tree Partition, Validation	571.25%

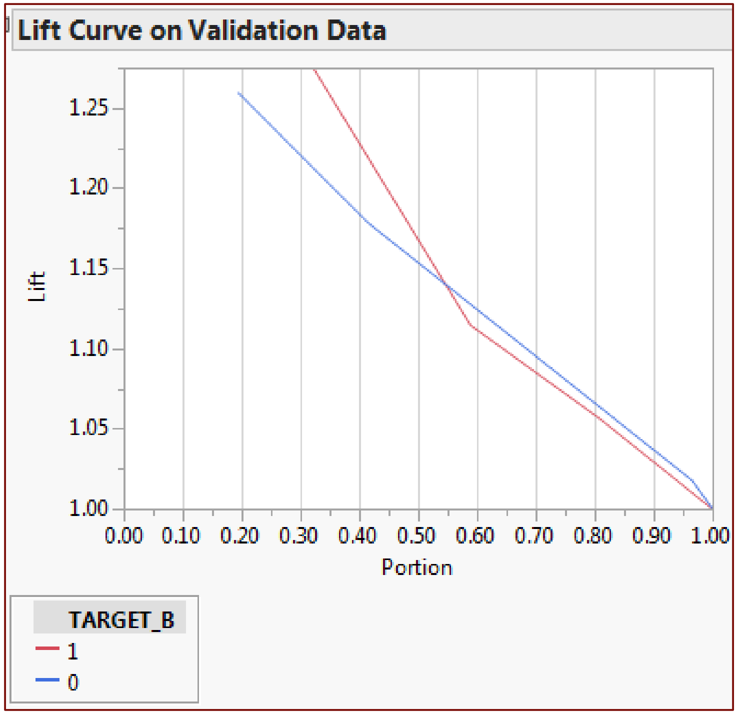
From here, I would favor the tree partition.

Lift Curves (Validation), Side-by-Side

Logistic Regression



Tree Partition

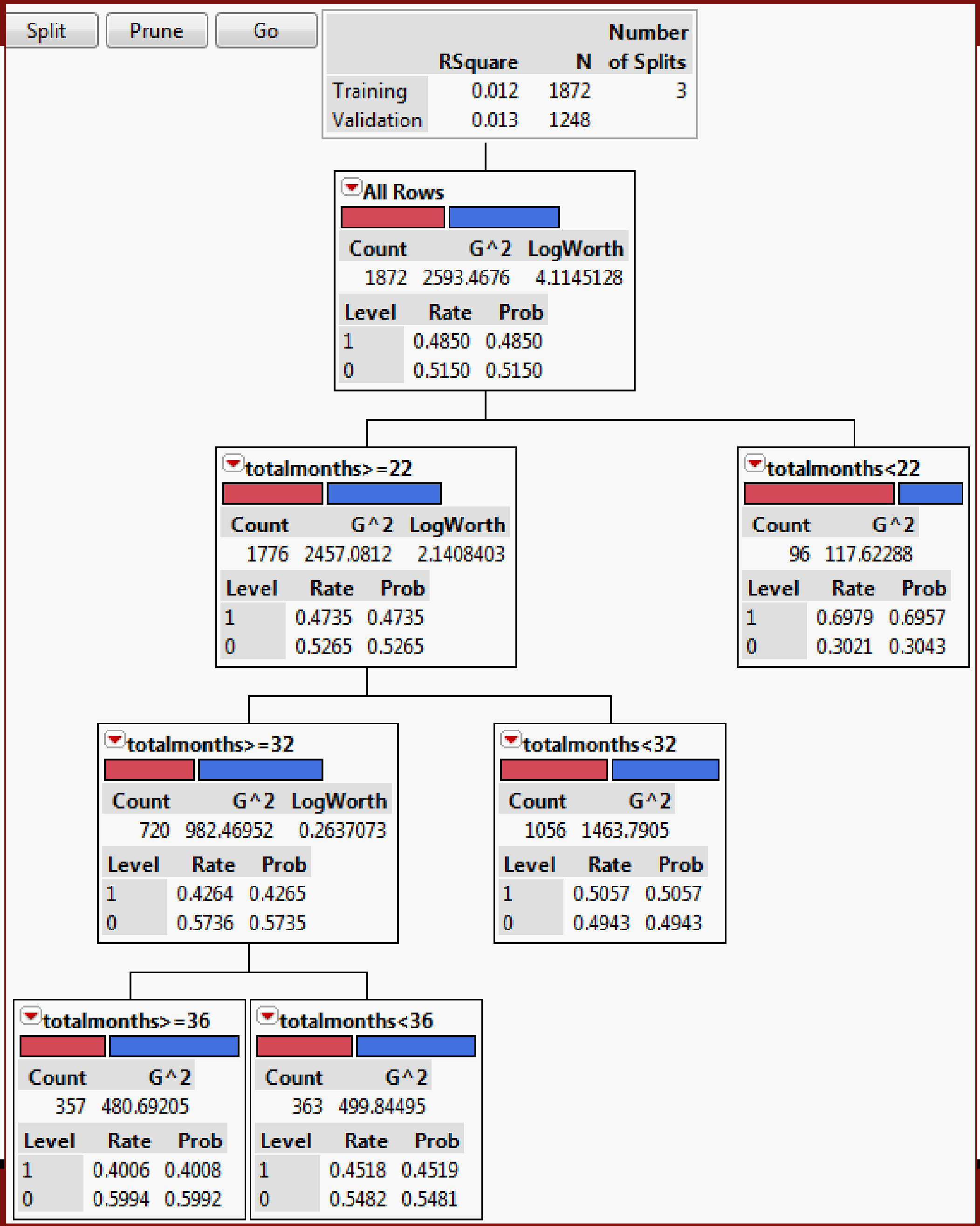


"When the response rate for a category is very low anyway (for example, a direct mail response rate), the lift curve explains things with more detail than the ROC curve." http://www.jmp.com/support/help/Graphs_for_Goodness_of_Fit.shtml

Best Model

- Tree Partition on *totalmonths* only

Months	Probability of Donation
0-21	69.57%
22-31	50.57%
32-35	45.19%
36+	40.08%



TESTING

New Incoming Data

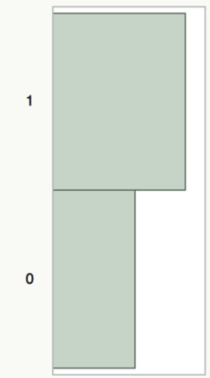
- I begin by assessing the probability of donation for each of the 2,000 new observations, according to our best model.
- I then classify those with 50% or greater probability as likely donors. 69 + 1,166 = 1,235
- JMP picks the same number of most likely donors. 1,235
- There are too many candidates to list on a single slide, as though I am preparing a mailing list.
- I have stuck the first 350, in descending order, in the Appendix. Jump to List

Row Id	totalmonths	Probability	Donor Bin
1	17	69.79%	1
2	33	45.19%	0
3	31	50.57%	1
4	31	50.57%	1
5	28	50.57%	1

	Probability	Count
Likely Donors	69.79%	69
	50.57%	1166
Not Likely Donors	45.19%	454
	40.08%	311

Distributions

Most Profitable Prediction for TARGET_B 2



Frequencies

Level	Count	Prob
0	765	0.38250
1	1235	0.61750
Total	2000	1.00000

N Missing 0

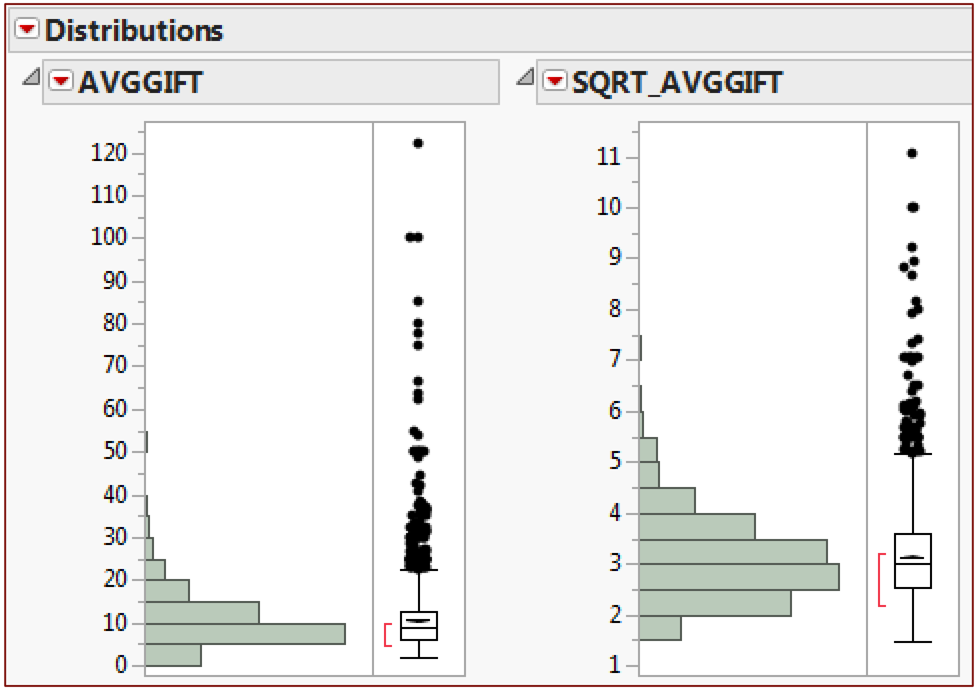
2 Levels

APPENDIX

JUMP TO RESULTS

Logistic Regression - Transformation

- In an effort to preserve the assumptions of ordinary least squares regression fits, I transformed some variables to reduce skew and outliers:



Logistic Regression - Interaction

- I also sought out possible interactions between variables.
- I wound up employing SQRT_AVGGIFT*gender_dummy in one of our regression models
- The idea to try that specific interaction came from an outside reference source:

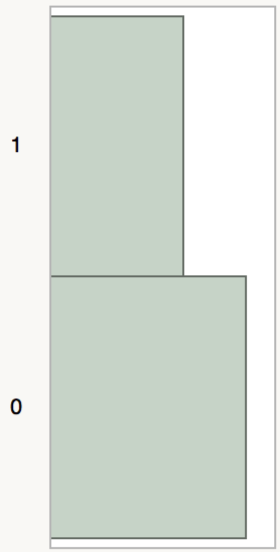
http://analytics.ncsu.edu/sesug/2008/MPSF-073.pdf

New Potential Donors, 1st 350, In Order

Row Id	totalmonths	Probability	Donor Bin
1	17	69.79%	1
26	17	69.79%	1
120	17	69.79%	1
165	17	69.79%	1
197	17	69.79%	1
199	17	69.79%	1
226	17	69.79%	1
306	17	69.79%	1
464	17	69.79%	1
719	17	69.79%	1
766	17	69.79%	1
792	17	69.79%	1
864	17	69.79%	1
1014	17	69.79%	1
1231	17	69.79%	1
1263	17	69.79%	1
1287	17	69.79%	1
1410	17	69.79%	1
1437	17	69.79%	1
1460	17	69.79%	1
1499	17	69.79%	1
1524	17	69.79%	1
1606	17	69.79%	1
1755	17	69.79%	1
1919	17	69.79%	1
6	18	69.79%	1
9	18	69.79%	1
17	18	69.79%	1
56	18	69.79%	1
223	18	69.79%	1
360	18	69.79%	1
403	18	69.79%	1
477	18	69.79%	1
722	18	69.79%	1
778	18	69.79%	1
833	18	69.79%	1
854	18	69.79%	1
1665	18	69.79%	1
1920	18	69.79%	1
1945	18	69.79%	1
1966	18	69.79%	1
215	19	69.79%	1
410	19	69.79%	1
489	19	69.79%	1
638	19	69.79%	1
648	19	69.79%	1
824	19	69.79%	1
1003	19	69.79%	1
1244	19	69.79%	1
1466	19	69.79%	1
1610	19	69.79%	1
1886	19	69.79%	1
36	20	69.79%	1
411	20	69.79%	1
646	20	69.79%	1
890	20	69.79%	1
936	20	69.79%	1
1596	20	69.79%	1
1626	20	69.79%	1
1641	20	69.79%	1
448	21	69.79%	1
938	21	69.79%	1
955	21	69.79%	1
960	21	69.79%	1
1002	21	69.79%	1
1119	21	69.79%	1
1175	21	69.79%	1
1233	21	69.79%	1
1363	21	69.79%	1
45	22	50.57%	1
735	22	50.57%	1
930	22	50.57%	1
937	22	50.57%	1
975	22	50.57%	1
999	22	50.57%	1
1199	22	50.57%	1
1684	22	50.57%	1
1809	22	50.57%	1
77	23	50.57%	1
80	23	50.57%	1
512	23	50.57%	1
741	23	50.57%	1
966	23	50.57%	1
1810	23	50.57%	1
412	24	50.57%	1
119	25	50.57%	1
326	25	50.57%	1
771	25	50.57%	1
409	26	50.57%	1
912	26	50.57%	1
1357	26	50.57%	1
1603	26	50.57%	1
1942	26	50.57%	1
1970	26	50.57%	1
1976	26	50.57%	1
340	27	50.57%	1
468	27	50.57%	1
1058	27	50.57%	1
1845	27	50.57%	1
5	28	50.57%	1
25	28	50.57%	1
50	28	50.57%	1
51	28	50.57%	1
69	28	50.57%	1
71	28	50.57%	1
72	28	50.57%	1
85	28	50.57%	1
90	28	50.57%	1
99	28	50.57%	1
101	28	50.57%	1
105	28	50.57%	1
135	28	50.57%	1
148	28	50.57%	1
162	28	50.57%	1
172	28	50.57%	1
174	28	50.57%	1
188	28	50.57%	1
214	28	50.57%	1
216	28	50.57%	1
254	28	50.57%	1
257	28	50.57%	1
263	28	50.57%	1
272	28	50.57%	1
276	28	50.57%	1
291	28	50.57%	1
305	28	50.57%	1
325	28	50.57%	1
329	28	50.57%	1
332	28	50.57%	1
338	28	50.57%	1
339	28	50.57%	1
346	28	50.57%	1
367	28	50.57%	1
375	28	50.57%	1
377	28	50.57%	1
380	28	50.57%	1
394	28	50.57%	1
420	28	50.57%	1
437	28	50.57%	1
438	28	50.57%	1
441	28	50.57%	1
449	28	50.57%	1
463	28	50.57%	1
465	28	50.57%	1
472	28	50.57%	1
485	28	50.57%	1
488	28	50.57%	1
498	28	50.57%	1
500	28	50.57%	1
505	28	50.57%	1
508	28	50.57%	1
535	28	50.57%	1
542	28	50.57%	1
547	28	50.57%	1
556	28	50.57%	1
566	28	50.57%	1
569	28	50.57%	1
575	28	50.57%	1
579	28	50.57%	1
582	28	50.57%	1
583	28	50.57%	1
593	28	50.57%	1
596	28	50.57%	1
606	28	50.57%	1
608	28	50.57%	1
609	28	50.57%	1
617	28	50.57%	1
619	28	50.57%	1
630	28	50.57%	1
639	28	50.57%	1
650	28	50.57%	1
651	28	50.57%	1
659	28	50.57%	1
676	28	50.57%	1
677	28	50.57%	1
695	28	50.57%	1
696	28	50.57%	1
708	28	50.57%	1
740	28	50.57%	1
742	28	50.57%	1
745	28	50.57%	1
746	28	50.57%	1
755	28	50.57%	1
759	28	50.57%	1
770	28	50.57%	1
780	28	50.57%	1
781	28	50.57%	1
783	28	50.57%	1
786	28	50.57%	1
793	28	50.57%	1
798	28	50.57%	1
803	28	50.57%	1
815	28	50.57%	1
817	28	50.57%	1
829	28	50.57%	1
840	28	50.57%	1
844	28	50.57%	1
861	28	50.57%	1
866	28	50.57%	1
870	28	50.57%	1
876	28	50.57%	1
879	28	50.57%	1
918	28	50.57%	1
932	28	50.57%	1
946	28	50.57%	1
953	28	50.57%	1
981	28	50.57%	1
987	28	50.57%	1
992	28	50.57%	1
1017	28	50.57%	1
1018	28	50.57%	1
1026	28	50.57%	1
1030	28	50.57%	1
1034	28	50.57%	1
1043	28	50.57%	1
1053	28	50.57%	1
1069	28	50.57%	1
1078	28	50.57%	1
1079	28	50.57%	1
1096	28	50.57%	1
1118	28	50.57%	1
1140	28	50.57%	1
1141	28	50.57%	1
1142	28	50.57%	1
1152	28	50.57%	1
1159	28	50.57%	1
1162	28	50.57%	1
1176	28	50.57%	1
1186	28	50.57%	1
1189	28	50.57%	1
1194	28	50.57%	1
1198	28	50.57%	1
1215	28	50.57%	1
1224	28	50.57%	1
1225	28	50.57%	1
1226	28	50.57%	1
1245	28	50.57%	1
1260	28	50.57%	1
1262	28	50.57%	1
1288	28	50.57%	1
1296	28	50.57%	1
1297	28	50.57%	1
1310	28	50.57%	1
1331	28	50.57%	1
1335	28	50.57%	1
1343	28	50.57%	1
1348	28	50.57%	1
1352	28	50.57%	1
1353	28	50.57%	1
1355	28	50.57%	1
1369	28	50.57%	1
1377	28	50.57%	1
1389	28	50.57%	1
1391	28	50.57%	1
1393	28	50.57%	1
1396	28	50.57%	1
1401	28	50.57%	1
1408	28	50.57%	1
1423	28	50.57%	1
1428	28	50.57%	1
1435	28	50.57%	1
1443	28	50.57%	1
1446	28	50.57%	1
1456	28	50.57%	1
1471	28	50.57%	1
1472	28	50.57%	1
1496	28	50.57%	1
1501	28	50.57%	1
1503	28	50.57%	1
1507	28	50.57%	1
1519	28	50.57%	1
1526	28	50.57%	1
1548	28	50.57%	1
1563	28	50.57%	1
1584	28	50.57%	1
1597	28	50.57%	1
1601	28	50.57%	1
1616	28	50.57%	1
1617	28	50.57%	1
1622	28	50.57%	1
1624	28	50.57%	1
1627	28	50.57%	1
1630	28	50.57%	1
1635	28	50.57%	1
1636	28	50.57%	1
1637	28	50.57%	1
1640	28	50.57%	1
1658	28	50.57%	1
1672	28	50.57%	1
1676	28	50.57%	1
1712	28	50.57%	1
1714	28	50.57%	1
1743	28	50.57%	1
1744	28	50.57%	1
1748	28	50.57%	1
1749	28	50.57%	1
1751	28	50.57%	1
1756	28	50.57%	1
1757	28	50.57%	1
1758	28	50.57%	1
1759	28	50.57%	1
1775	28	50.57%	1
1787	28	50.57%	1
1801	28	50.57%	1
1807	28	50.57%	1
1822	28	50.57%	1
1826	28	50.57%	1
1839	28	50.57%	1
1857	28	50.57%	1
1883	28	50.57%	1
1892	28	50.57%	1
1914	28	50.57%	1
1936	28	50.57%	1
1947	28	50.57%	1
1956	28	50.57%	1
1960	28	50.57%	1
1961	28	50.57%	1
1975	28	50.57%	1
1992	28	50.57%	1
32	29	50.57%	1
38	29	50.57%	1
41	29	50.57%	1
65	29	50.57%	1
78	29	50.57%	1
81	29	50.57%	1
95	29	50.57%	1
100	29	50.57%	1
103	29	50.57%	1
107	29	50.57%	1
113	29	50.57%	1
130	29	50.57%	1
169	29	50.57%	1
186	29	50.57%	1
187	29	50.57%	1
206	29	50.57%	1
209	29	50.57%	1
227	29	50.57%	1
246	29	50.57%	1
247	29	50.57%	1
256	29	50.57%	1
258	29	50.57%	1
271	29	50.57%	1
277	29	50.57%	1
279	29	50.57%	1
283	29	50.57%	1
307	29	50.57%	1
317	29	50.57%	1
319	29	50.57%	1
321	29	50.57%	1

Distributions

Most Likely TARGET_B



Frequencies

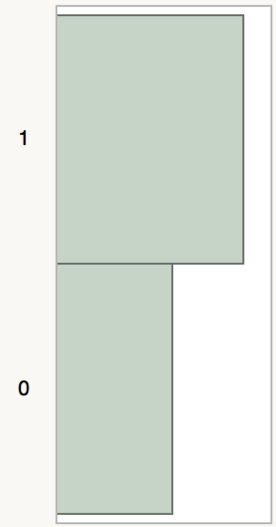
Level	Count	Prob
0	1194	0.59700
1	806	0.40300
Total	2000	1.00000

N Missing 0
2 Levels

Classification count for the new data using Linear Regression

Distributions

Most Profitable Prediction for TARGET_B 2



Frequencies

Level	Count	Prob
0	765	0.38250
1	1235	0.61750
Total	2000	1.00000

N Missing 0
2 Levels

Classification count for new data using Decision Tree