Topic 6. Interactions

Case 3: Donor Recapture

using Transaction, Overlay, and Census Data

The Plan

The decision tree analysis suggested some possible improvements to our regression model that we shall investigate:

1. Pruning to eliminate some STATE dummies.

2. Pruning to eliminate MAILCODE.

3. Adding a LASTGIFT by PEPSTRFL interaction.

4. Replacing STATE dummies by longitude and latitude – another stab at pruning.

Where We Are

The regression model is preferred, but there are suggestions from the decision tree fits that it can be improved.

Here are the features in the regression model ...

Table 17. Features in Regression

File	Feature	Type	Number of Dummies
464	LASTGIFT	num	
75	PEPSTRFL	chr	1
4	STATE	chr	31
11	RECP3	chr	1
8	DOB	num	
6	MAILCODE	chr	1
359	MHUC2	num	
465	LASTDATE	num	
460	MINRAMNT	num	

Table 18. Feature Definitions

File	Feature	Type	Definition
464	LASTGIFT	num	Dollar amount of most recent gift
75	PEPSTRFL	chr	Has given to three consecutive card mailings
4	STATE	chr	State of residence
11	RECP3	chr	Has given to CTY's P3 program
8	DOB	num	Date of birth
6	MAILCODE	chr	Mailing address is correct
359	MHUC2	num	Census tract homeowner cost w/out mortgage
465	LASTDATE	num	Date associated with the most recent gift
460	MINRAMNT	num	Dollar amount of smallest gift to date

Tree Suggestions

The tree suggested

1. Not all the state dummies should be in the model.
2. A LASTGIFT by PEPSTRFL interaction (definition repeated a few slides from now) should be in the model.
3. MAILCODE should not be in the model

Let's see if the regression coefficients tell the same tale ...

Table 19. Regression Coefficients
(regr/cty_lif.r.Rout)

Feature	Estimate	Std. Error	t-value	p-value	
Intercept	-1.911e+01	3.402e+00	-5.616	1.96e-08	***
LASTGIFT	2.126e-02	1.415e-03	15.026	$< 2e-16$	***
PEPSTRFL	2.010e-01	3.749e-02	5.361	8.30e-08	***
STATE.AR	6.236e-02	2.127e-01	0.293	0.769425	
STATE.AZ	1.399e-01	1.700e-01	0.823	0.410770	
STATE.CA	4.137e-01	1.366e-01	3.028	0.002467	**
STATE.CO	2.686e-01	1.772e-01	1.516	0.129520	
STATE.FL	1.584e-01	1.430e-01	1.107	0.268117	
STATE.GA	1.496e-01	1.588e-01	0.942	0.346416	
STATE.HI	6.331e-01	2.819e-01	2.246	0.024709	*
STATE.IA	-2.583e-02	1.969e-01	-0.131	0.895609	
STATE.ID	5.674e-01	2.632e-01	2.156	0.031089	*
STATE.IL	1.049e-01	1.471e-01	0.713	0.476022	
STATE.IN	1.351e-02	1.629e-01	0.083	0.933913	
STATE.KS	7.852e-02	1.980e-01	0.397	0.691704	
STATE.KY	7.040e-02	1.864e-01	0.378	0.705622	
STATE.LA	4.164e-02	1.865e-01	0.223	0.823344	
STATE.MI	1.088e-01	1.492e-01	0.729	0.465977	
STATE.MN	-7.062e-02	1.730e-01	-0.408	0.683147	
STATE.MO	1.375e-01	1.655e-01	0.831	0.405837	
STATE.MT	1.700e-01	2.681e-01	0.634	0.526118	

Table 19. Regression Coefficients (continued)

Feature	Estimate	Std. Error	t-value	p-value	
STATE.NC	1.928e-01	1.544e-01	1.249	0.211818	
STATE.NM	2.802e-01	2.231e-01	1.256	0.209036	
STATE.NV	7.349e-02	2.180e-01	0.337	0.735998	
STATE.OK	-7.238e-02	1.855e-01	-0.390	0.696369	
STATE.OR	4.053e-01	1.724e-01	2.350	0.018754	*
STATE.S1	-5.482e-01	4.914e-01	-1.116	0.264629	
STATE.S2	3.605e-01	2.438e-01	1.479	0.139193	
STATE.S3	-1.031e-01	1.810e-01	-0.569	0.569026	
STATE.S4	1.207e-01	2.128e-01	0.568	0.570346	
STATE.S5	2.643e-01	1.749e-01	1.512	0.130650	
STATE.TN	1.720e-02	1.683e-01	0.102	0.918618	
STATE.TX	1.236e-01	1.446e-01	0.855	0.392714	
STATE.WA	2.685e-01	1.583e-01	1.696	0.089902	.
STATE.WI	-6.061e-02	1.654e-01	-0.366	0.714104	
RECP3	5.765e-01	1.226e-01	4.702	2.58e-06	***
DOB.miss	-2.427e-01	9.665e-02	-2.511	0.012035	*
DOB.linr	1.969e-04	5.209e-05	3.780	0.000157	***
DOB.quad	-2.615e-08	6.795e-09	-3.849	0.000119	***
MAILCODE	-4.274e-01	1.452e-01	-2.942	0.003259	**
MHUC2	5.505e-02	2.060e-02	2.673	0.007525	**
LASTDATE	2.002e-03	3.560e-04	5.624	1.88e-08	***
MINRAMNT	-4.458e-03	2.414e-03	-1.847	0.064729	.

What the Coefficients Say

Our previous crude slope estimates look about right: All else equal, one can expect $0.02 for every $1 of LASTGIFT. Presumably our interaction hypothesis will pan out.

Judging from the t-statistics, some STATE dummies do look superfluous.

It looks like MAILCODE should stay. One would think that a bad address means no gift but apparently mail gets forwarded. There were 965 bad addresses in the learning sample of which 29 gave gifts; the average of these 29 was $13.14.

State Dummies

Recall that we used upward selection based on $mse.val$ for our model but we put all dummies for a variable in as a lump.

We shall prune by taking them out one at a time for as long as $mse.val$ declines ...

Pruning Output

charity/prun/cty_prun.r.Rout

```
HI deleted from X
WI deleted from X
NC deleted from X
OK deleted from X
GA deleted from X
MN deleted from X
S2 deleted from X
TN deleted from X
ID deleted from X
IN deleted from X
IL deleted from X
MO deleted from X
AR deleted from X
```


Table 20. Performance Measures

Model	Specification	Mean Squared Error Learning	Validation	Test
Mean	learning sample	20.09922	18.82322	17.86605
Regr	selected model	19.96083	18.67709	17.80003
Regr	pruned model	19.96681	18.66845	17.81442
Nnet	6 iter X 5 HU	19.97731	18.72594	17.85258
Tree	$cp = 0.001$	19.89110	18.88466	18.07888
Tree	$cp = 0.0008$	19.80992	18.83118	18.31281
Tree	$cp = 0.0001$	19.01715	19.64272	18.90903

Pruning

Pruning did not help much.

We'll just leave all the STATE dummies in the model and move on.

Let's see how the suggestion of an interaction between LASTGIFT and PEPSTRFL pans out ...

Interactions

An interaction is the derived feature obtained by multiplying LASTGIFT and PEPSTRFL:

> LAST.by.PEP = LASTGIFT x PEP-STRFL

Because PEPSTRFL is a dummy variable, this has the effect of fitting two linear functions of LASTGIFT to the data: There is one slope coefficient for three-time givers (PEP-STRFL=1) and a different slope for others (PEPSTRFL=0).

The slope for others is the coefficient of LAST-GIFT. The slope for three-time givers is the sum of the coefficients of LASTGIFT and LAST.by.PEP.

Table 21. Performance Measures

		Mean Squared Error		
Model	Specification	Learning	Validation	Test
Mean	learning sample	20.09922	18.82322	17.86605
Regr	selected model	19.96083	18.67709	17.80003
Regr	selected + inter	19.91118	18.60526	17.79515
Nnet	6 iter X 5 HU	19.97731	18.72594	17.85258
Tree	$cp = 0.001$	19.89110	18.88466	18.07888
Tree	$cp = 0.0008$	19.80992	18.83118	18.31281
Tree	$cp = 0.0001$	19.01715	19.64272	18.90903

Success

The interaction improved the model substantially.

The interaction stays!

Let's see what the coefficients look like now ...

Table 22. Regression Coefficients
(intr/cty_intr.r.Rout)

Feature	Estimate	Std. Error	t-value	p-value	
Intercept	-1.976e+01	3.398e+00	-5.816	6.07e-09	***
LASTGIFT	7.567e-03	1.767e-03	4.283	1.85e-05	***
PEPSTRFL	-3.302e-01	5.562e-02	-5.936	2.93e-09	***
LAST.by.PEP	3.472e-02	2.689e-03	12.912	$<$ 2e-16	***
STATE.AR	5.091e-02	2.125e-01	0.240	0.810633	
STATE.AZ	1.390e-01	1.698e-01	0.819	0.413017	
STATE.CA	4.036e-01	1.365e-01	2.958	0.003101	**
STATE.CO	2.650e-01	1.769e-01	1.498	0.134211	
STATE.FL	1.566e-01	1.429e-01	1.096	0.272890	
STATE.GA	1.465e-01	1.586e-01	0.924	0.355617	
STATE.HI	6.318e-01	2.815e-01	2.244	0.024823	*
STATE.IA	-1.748e-02	1.966e-01	-0.089	0.929180	
STATE.ID	5.446e-01	2.629e-01	2.072	0.038293	*
STATE.IL	1.017e-01	1.469e-01	0.692	0.489065	
STATE.IN	1.430e-02	1.627e-01	0.088	0.929992	
STATE.KS	7.198e-02	1.978e-01	0.364	0.715886	
STATE.KY	7.201e-02	1.861e-01	0.387	0.698835	
STATE.LA	4.555e-02	1.863e-01	0.245	0.806829	
STATE.MI	1.112e-01	1.490e-01	0.746	0.455621	
STATE.MN	-6.804e-02	1.728e-01	-0.394	0.693752	
STATE.MO	1.503e-01	1.653e-01	0.910	0.363041	
STATE.MT	1.683e-01	2.678e-01	0.628	0.529690	

Table 22. Regression Coefficients (continued)

Feature	Estimate	Std. Error	t-value	p-value	
STATE.NC	1.936e-01	1.543e-01	1.255	0.209558	
STATE.NM	2.740e-01	2.228e-01	1.230	0.218841	
STATE.NV	7.333e-02	2.177e-01	0.337	0.736222	
STATE.OK	-7.684e-02	1.852e-01	-0.415	0.678292	
STATE.OR	3.836e-01	1.722e-01	2.227	0.025943	*
STATE.S1	-5.454e-01	4.908e-01	-1.111	0.266505	
STATE.S2	3.526e-01	2.435e-01	1.448	0.147487	
STATE.S3	-1.053e-01	1.808e-01	-0.583	0.560198	
STATE.S4	1.296e-01	2.125e-01	0.610	0.542080	
STATE.S5	2.649e-01	1.747e-01	1.517	0.129351	
STATE.TN	1.160e-02	1.681e-01	0.069	0.944966	
STATE.TX	1.185e-01	1.444e-01	0.821	0.411659	
STATE.WA	2.643e-01	1.581e-01	1.671	0.094651	.
STATE.WI	-4.792e-02	1.652e-01	-0.290	0.771823	
RECP3	6.015e-01	1.225e-01	4.912	9.03e-07	***
DOB.miss	-2.269e-01	9.654e-02	-2.350	0.018756	*
DOB.linr	1.906e-04	5.203e-05	3.664	0.000248	***
DOB.quad	-2.576e-08	6.786e-09	-3.796	0.000147	***
MAILCODE	-4.372e-01	1.451e-01	-3.014	0.002579	**
MHUC2	5.183e-02	2.057e-02	2.520	0.011753	*
LASTDATE	2.091e-03	3.557e-04	5.880	4.13e-09	***
MINRAMNT	3.675e-03	2.492e-03	1.475	0.140218	

What the Coefficients Say

Recall that the slope for others is the coefficient of LASTGIFT. The slope for three-time givers is the sum of the coefficients of LASTGIFT and LAST.by.PEP.

Thus, all else being equal we can expect someone who is not a three-time giver to give \$0.008 per dollar of LASTGIFT whereas we can expect a three-time give to give \$0.042 per dollar of LASTGIFT.

These slopes are reasonably consistent with the estimates in Table 15, which were \$0.014 and \$0.028 per dollar, respectively.

STATE Dummies Revisited

STATE is just an indicator of location. Perhaps the exact location given by an address's longitude and latitude would work better.

There are databases that give the longitude and latitude of every ZIP code: See, for example,
http://www.zipinfo.com/search/zipcode.htm

Derived Features

We shall replace the STATE dummies by eight variables:

LON, LONSQR, LAT, LATSQR, LONLAT, HI, AK, TERRITORY

- If the ZIP code is in one of the 48 states the longitude and latitude variables are filled in from the database and HI, AK, TERRITORY are zero.
- If the ZIP code is in Hawaii the longitude and latitude variables are zero, HI is one, AK is zero, and TERRITORY is zero.
- If the ZIP code is in ALASKA the longitude and latitude variables are zero, AK is one, HI is zero, and TERRITORY is zero.
- Otherwise the longitude and latitude variables are zero, TERRITORY is one, HI is zero, and AK is zero.

Table 23. Performance Measures

		Mean Squared Error		
Model	Specification	Learning	Validation	Test
Mean	learning sample	20.09922	18.82322	17.86605
Regr	selected model	19.96083	18.67709	17.80003
Regr	selected + inter	19.91118	18.60526	17.79515
Regr	lon, lat, + inter	19.91875	18.60238	17.79226
Nnet	6 iter X 5 HU	19.97731	18.72594	17.85258
Tree	$cp = 0.001$	19.89110	18.88466	18.07888
Tree	$cp = 0.0008$	19.80992	18.83118	18.31281
Tree	$cp = 0.0001$	19.01715	19.64272	18.90903

Similar to Pruning

Geographical location works slightly better than STATE dummies but not better than pruned STATE dummies.

Note that learning MSE went up and validation MSE went down, which means that we got rid of some overfitting, which is what pruning did.

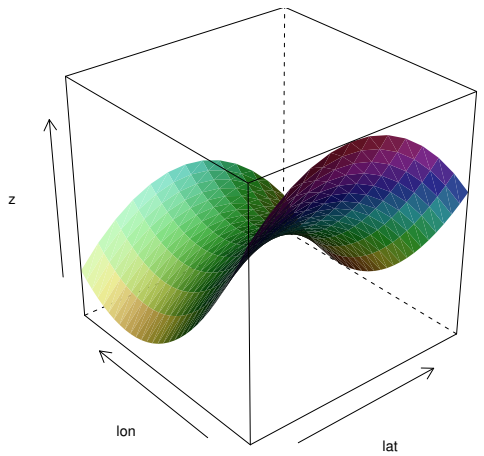
The coefficients are interesting, however ...

Table 24. Regression Coefficients

(usa/cty_usa.r.Rout)

Feature	Estimate	Std. Error	t-value	p-value	
Intercept	-1.948e+01	3.419e+00	-5.697	1.22e-08	***
LASTGIFT	7.585e-03	1.767e-03	4.293	1.76e-05	***
PEPSTRFL	-3.310e-01	5.560e-02	-5.954	2.63e-09	***
LON	3.689e-02	1.479e-02	2.494	0.012643	*
LONSQR	2.273e-04	8.765e-05	2.594	0.009501	**
LAT	8.273e-02	3.697e-02	2.238	0.025232	*
LATSQR	-1.108e-03	6.815e-04	-1.626	0.103866	
LONLAT	5.561e-05	3.135e-04	0.177	0.859209	
HI	3.925e-01	4.744e-01	0.827	0.407984	
AK	-4.635e-01	5.133e-01	-0.903	0.366626	
TERRITORY	6.006e-01	6.217e-01	0.966	0.334015	
RECP3	6.033e-01	1.224e-01	4.927	8.36e-07	***
DOB_miss	-2.166e-01	9.623e-02	-2.251	0.024361	*
DOB_linr	1.882e-04	5.198e-05	3.621	0.000293	***
DOB_quad	-2.552e-08	6.782e-09	-3.762	0.000169	***
MAILCODE	-4.341e-01	1.450e-01	-2.993	0.002762	**
MHUC2	5.983e-02	2.036e-02	2.939	0.003299	**
LASTDATE	2.084e-03	3.556e-04	5.862	4.60e-09	***
MINRAMNT	3.760e-03	2.491e-03	1.509	0.131239	
LAST.by.PEP	3.474e-02	2.689e-03	12.920	¡ 2e-16	***

Fig 70. Longitude and Latitude Effects



Shown is the surface implied by the regression coefficients LON, LONSQR, LAT, LATSQR, LONLAT shown in Table 24 for longitude and latitude ranging over the lower 48 states.

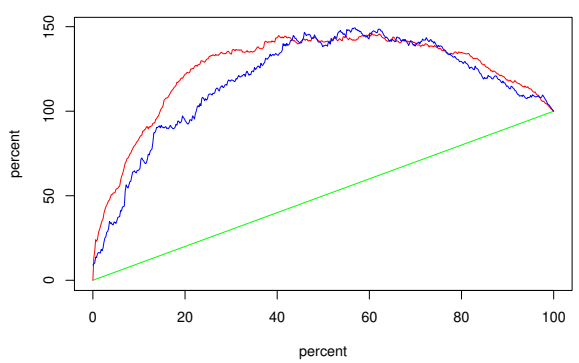
Lift Charts

We're done!

We'll use the selected model+interactions specification.

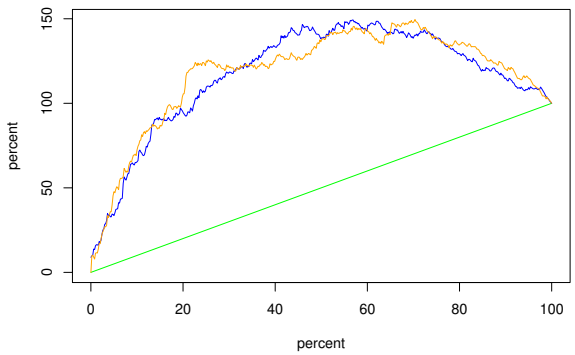
Let's have one last look at some lift charts ...

Fig 71. Lift Charts



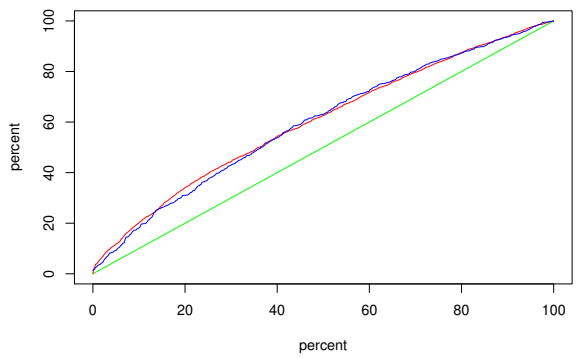
The green curve shows net revenue if persons in the learning sample were mailed solicitations in random order. The red curve shows net revenue in the learning sample if persons are sorted by their predicted gift and mailed solicitations in sorted order, highest first; blue is the same for the validation sample. The plots are normalized so endpoints plot at (100,100). Net revenue is the gift less a mailing cost of $0.68.

Fig 72. Lift Charts



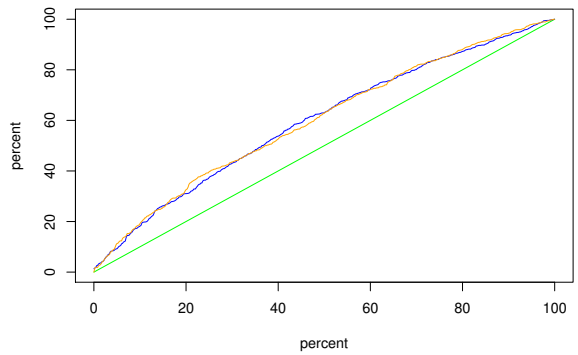
Same as Fig 71 except that the orange line is the blue line from Fig 54, which shows the lift of the regression model in the validation sample.

Fig 73. Conventional Lift Charts



Same as Fig 71 but gross revenue instead of net revenue.

Fig 74. Conventional Lift Charts



Same as Fig 73 except that the orange line is the blue line from Fig 55, which shows the lift of the regression model in the validation sample.

Main Point

Adding the interaction of LASTGIFT and PEP-STRFL,

LAST.by.PEP = LASTGIFT x PEP-STRFL

to the selected model substantially improved results.

This discovery illustrates the fact that one of the main attractions of decision trees is their ability to identify potentially useful features and derived features.

Blank page