

DSC 433/533 – Class 6 – Charity Case Example

Task: A major charity wishes to solicit a donation. Which donor characteristics are associated with large donations? Can we use this information to target the solicitations.

25 measured characteristics (features) pre-selected from an original database with 481 features.

What to do with nominal features?

STATE: 4 groups coded with 3 indicator variables, STGP_2, STGP_3, and STGP_4.

RFA: 4 groups coded with 3 indicator variables, RFA_2A_E, RFA_2A_F, and RFA_2A_G.

1176 donors responded to the last mailing (original database contains 95412 former donors):
588 (50%) *training*, 353 (30%) *validation*, 235 (20%) *test* (seed: 12345).

Consider predicting the donation amount for each donor.

Multiple linear regression model with variable selection

- Select XLMiner > Prediction > Multiple Linear Regression.
- In step 1 move all the variables except STATE, STGP, AGE, RFA_2, and RFA_2A to the “Input variables” box and “TARGET_D” to the “Output variable” box.
- In step 2 click on “Best subset” and experiment with each of the 5 methods.
- Select “Summary report” for “Score training data.”
- Select “Detailed report,” “Summary report,” and “Lift charts” for “Score validation data.”
- De-select “Summary report” for “Score test data” (we won’t be using test data for this example).

Rules of thumb:

- Consider first model with $C_p < \#Coeffs (= \#Predictors + 1)$ and all subsequent models until C_p starts to go up.
- Consider first model with $C_p < \#Coeffs (= \#Predictors + 1)$ and all subsequent models until Adjusted R-squared starts to go down.

Backward:

#Coeffs	C_p	Adj. R-Squared	Model (Constant present in all models)									
			1	2	3	4	5	6	7	8	9	
5	-0.12	0.764	Constant	MAXRAMNT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*	*	*	*
6	-0.09	0.765	Constant	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*	*	*
7	-2.73	0.766	Constant	NUMPRM12	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*	*
8	-2.00	0.766	Constant	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*
9	-0.99	0.766	Constant		HV2	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G
10	-0.54	0.766	Constant		HV2	IC4	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F

Forward:

#Coeffs	Cp	Adj. R-Squared	Model (Constant present in all models)								
			1	2	3	4	5	6	7	8	9
6	0.26	0.764	Constant	MAXRAMNT	AVGGIFT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*	*
7	0.68	0.765	Constant	MINRAMNT	MAXRAMNT	AVGGIFT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*
8	1.14	0.765	Constant	MINRAMNT	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*
9	-1.14	0.766	Constant	NUMPRM12	MINRAMNT	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G
10	-0.33	0.766	Constant	MHUC1	NUMPRM12	MINRAMNT	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F
11	0.33	0.766	Constant	IC4	MHUC1	NUMPRM12	MINRAMNT	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE

Exhaustive:

#Coeffs	Cp	Adj. R-Squared	Model (Constant present in all models)								
			1	2	3	4	5	6	7	8	9
5	-0.12	0.764	Constant	MAXRAMNT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*	*	*
6	-0.09	0.765	Constant	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*	*
7	-2.73	0.766	Constant	NUMPRM12	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*
8	-2.00	0.766	Constant	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*
9	-1.24	0.766	Constant	IC4	MHUC1	NUMPRM12	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G
10	-0.68	0.766	Constant	IC4	MHUC1	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F

Sequential:

#Coeffs	Cp	Adj. R-Squared	Model (Constant present in all models)								
			1	2	3	4	5	6	7	8	9
5	-0.12	0.764	Constant	MAXRAMNT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*	*	*
6	0.26	0.764	Constant	MAXRAMNT	AVGGIFT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*	*
7	2.20	0.764	Constant	STGP_2	MAXRAMNT	AVGGIFT	LASTGIFT	RFA_2A_F	RFA_2A_G	*	*
8	3.20	0.764	Constant	STGP_2	STGP_3	MAXRAMNT	AVGGIFT	LASTGIFT	RFA_2A_F	RFA_2A_G	*
9	-1.24	0.766	Constant	IC4	MHUC1	NUMPRM12	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G
10	-0.68	0.766	Constant	IC4	MHUC1	NUMPRM12	MAXRAMNT	AVGGIFT	LASTGIFT	LASTDATE	RFA_2A_F

Stepwise (3, 2.99):

#Coeffs	Cp	Adj. R-Squared	Model (Constant present in all models)								
			1	2	3	4	5	6	7	8	9
6	-0.09	0.765	Constant	MAXRAMNT	LASTGIFT	LASTDATE	RFA_2A_F	RFA_2A_G	*	*	*

Model	RMSE in validation sample	Lift in first decile
No predictors	10.97	1.000
All predictors (except AGE)	6.24	2.235
All predictors + AGE	6.24	2.235
Log-transform response	6.64	2.207
Subset with 4 predictors	6.21	2.254
Subset with 5 predictors	6.20	2.267
Subset with 6 predictors	6.25	2.267
Subset with 7 predictors	6.24	2.248
Subset with 8 predictors	6.25	2.273
Subset with 9 predictors	6.24	2.254