

## DSC 433/533 – Class 1 – XLMiner Example (class1.xls)

Everyone takes a slip of paper: 20 training (middle), 12 validation (side), 8 test (side).

**Task:** outdoor equipment retailer with a large customer database (80k) intends to mail 20k new catalogs to a quarter of its customers at a cost of \$2 per catalog (production/mailing costs). How much profit does the company expect to make?

Sample mailing: 40 customers (20 to *train* models, 12 to *validate* models, 8 to *test* models). The 8 “test” customers spent a total of \$139, or \$17.375 per customer.

Expected profit if randomly select 20k customers to mail =  $(17.375 - 2) * 20,000 = \$307,000$ .

Can we do better by using data on customer characteristics?

- recency (months since last purchase, -)
- frequency (number of purchases over the last five years, +)
- average amount (spent per purchase, +)
- address (1 if urban, 0 if rural, +)

Idea: use sample data to order customers by predicted amount of spending and use results to pick off the top 25% of customers to send catalog to.

### Classification Step

Model probability of success (purch = 1) using logistic regression or classification trees.

- Logistic regression

Use training data to fit model (estimate probability equation):

$$\text{Prob} = 1 / (1 + \exp(2.53 + 0.21\text{rec} - 0.15\text{freq} - 0.14\text{avg} - 0.30\text{urb}))$$

e.g., id 21:  $\text{Prob} = 1 / (1 + \exp(2.53 + 0.21(17) - 0.15(15) - 0.14(10) - 0.30(0))) = 0.08$  (0)

e.g., id 28:  $\text{Prob} = 1 / (1 + \exp(2.53 + 0.21(5) - 0.15(12) - 0.14(16) - 0.30(1))) = 0.68$  (1)

Use validation data to assess accuracy of fitted model:

If the decision rule is “Predict success if Prob > 0.5” then classification matrix is:

	Predicted Class	
Actual Class	1	0
1	5	0
0	2	5

In other words, 2 false positives but no false negatives.

- Classification tree

Use training data to fit model (estimate tree rules):

Predict success if  $\text{rec} \leq 6.5$  and failure if  $\text{rec} > 6.5$ .

e.g., id 22:  $\text{rec} = 24$  so predict purch = 0 (0)

e.g., id 29:  $\text{rec} = 6$  so predict purch = 1 (1)

Use validation data to assess accuracy of fitted model:

	Predicted Class	
Actual Class	1	0
1	3	2
0	1	6

In other words, 1 false positive and 2 false negatives.

Conclusion: choose logistic regression model.

### Prediction Step

Model amount of spending using multiple linear regression or regression trees.

- Multiple linear regression

Use training data (9 cases with spend>0) to fit model (estimate regression equation):

$$E(\text{spend}) = -4.62 - 1.15 \text{ rec} + 1.32 \text{ freq} + 0.76 \text{ avg} + 13.45 \text{ urban}$$

e.g., id 36:  $E(\text{spend}) = -4.62 - 1.15 (9) + 1.32 (6) + 0.76 (40) + 13.45 (1) = 36.8$  (spend = 42)

Use validation data to assess accuracy of fitted model: root mean square error = 5.26.

- Regression tree

Use training data (9 cases with spend>0) to fit model (estimate tree rules):

$$E(\text{spend}) = 25.25 \text{ if } 17 < \text{avg} \leq 30 \text{ \& rec} \leq 19$$

$$E(\text{spend}) = 52 \text{ if } \text{avg} > 30 \text{ \& rec} \leq 7$$

$$E(\text{spend}) = 35 \text{ if } \text{avg} > 30 \text{ \& rec} > 11.5$$

e.g., id 38: avg = 37 & rec = 14 so  $E(\text{spend}) = 35$  (spend = 35)

Use validation data to assess accuracy of fitted model: root mean square error = 9.05.

Conclusion: choose multiple linear regression model.

### Test models

Use test data to assess performance of selected models.

- Predict purchase probabilities using logistic regression model (A).
- Predict spending for purchasers using multiple linear regression model (B).
- Multiply A and B together to get expected spending for everyone (C).
- Sort 8 test customers in order of C (high to low).
- Send offers to top 25% prospects (i.e. id's 8 and 7).
- e.g., id 8 expected spending =  $0.964 * 44.150 = 42.545$ .
- Lift =  $\frac{\text{cumulative expected model spending}}{\text{cumulative expected random spending}} = \frac{54 + 42}{17.375 + 17.375} = 2.763$
- Expected profit if send 20k catalogs to top 25% prospects =  $(2.763 * 17.375 - 2) * 20,000 = (48 - 2) * 20,000 = \$920,000$ .