PRED 450 55 Spring 2016
Solo 3: XYZ Company Database Campaign Response Modeling
v 1

Your Mission

Use XYZ's data to develop a model-based method for targeting customers that XYZ can use to make its upcoming mail campaign profitable. This campaign will consist of one mailing to XYZ customers who are selected based on your analytic results and your recommendations.

(1) Develop the best predictive model you can using initial responses (i.e., binary responses to first mailing) to XYZ's most recent mail campaign (the 16th one) *as your dependent variable. You may consider and apply either Frequentist or Bayesian predictive modeling methods, or both. Assess different models, and compare them using statistical methods to select your best one.*

A "good" model will have predictive accuracy, and the more of it, the better. It will help inform XYZ about which of its customers to mail to. It might even suggest things that XYZ can to to increase the likelihood of sales by changing its marketing mix.

(2) Assume that it costs XYZ $1.00 to produce and mail a catalogue to a customer, and that XYZ has recently earned a 10% profit on items it sells through its catalogues, excluding the costs of the catalogue and mailing.

(3) Using your model results and the other information available, identify which of XYZ's customers should be sent a catalogue, and what net revenue (net profit from sales less mailing costs) you expect will result from mailing to them. Include in your report how you did your calculations.

(4) "Score" (see below) customers that XYZ didn't target in its most recent campaign so that XYZ can mail to those you predict will produce net revenue as a result. Describe your method and include your estimate of expected net revenue from mailing to these customers. Be sure to mention any key assumptions or limitations.

(5) Compare the net revenue you expect based on your modeling to the revenue that actually resulted from XYZ's 16th campaign, and explain any difference you observe.

(6) Interpret the results of your most preferred model from a marketing perspective so that XYZ can learn about its customers.

(7) Address how XYZ can apply your final model to new customers*. XYZ would like to be able to predict responses to their next campaign by customers who are yet to be acquired by XYZ.*

(8) Suppose that it would be possible to assess your model's ability to predict responses to XYZ's next (The 17th) campaign by targeting selected customers. How would you design this "test?" What comparisons would you make?

XYZ is counting on you to help make their next campaign as lucrative as possible.

Be sure to succinctly summarize your methodology, and also the basis(es) for the various decisions you made, and the conclusions you drew. Be sure to address all objectives and

issues, and to qualify your results by describing any limitations of your methodology.

Be sure to see **Guidance**, below.

Materials

This assignment involves working with XYZ's database, which is more complicated than the data sets you used previously in this course. On Canvas you'll find some R data files, some documents describing the data, and some supporting materials. There are also videos about the creating the analytic data set and about evaluating targeting models in the Connect Recordings.

Working with the Data

Although there is documentation on XYZ's data, it is neither complete nor completely correct in all cases. You'll need to make the best use of the available information about the data that you can, as you would if you were doing a "real" project.

Prof. Miller has done much of the slog work in putting an analytic data set together for you. To fully understand the variables he has created for your use, you should review his R code, which although lengthy is pretty well documented. You'll find it in the file miller_xyz_data_integration.r. Open this file in an editor and review his code and comments. You'll also want to be sure to view the Connect recording he made describing how he constructed this combined data set. The recording is called "Special Solitary Sync Session to Review XYZ data work."

If you happen to take PRED 420, you'll run into this XYZ data, or at least versions of it, again.

Guidance

This is a target marketing response modeling exercise. Your task is to develop a model that predicts responses to the last (the 16th) mail campaign using the data you have available, and then to use it to tell XYZ which customers it should send a catalogue to in its next (17^{th}) mail campaign so that the net returns (total returns less mailing costs) realized will be as large as possible.

You can use any of a number of different kinds of classifier or binary limited dependent variable regression models to do this kind of response modeling. For summarizing models and comparing them, you'll want to consider using ROC curves and lift charts, so see R packages like gains, ROCR, AUC, Epi, and BCA. (A question you might want to consider is, how can you tell whether the AUCs of two models differ reliably, i.e. "significantly?" How about whether an AUC is reliably greater than 0.5?)

Many would start out on on a task like this by fitting a binary logistic or probit regression

model. Either may suffice for this assignment if its performance is adequate. But you could (also) use other methods as well, like cluster-wise regression or CART, an ensemble technique like Random Forest, or a margin classifier like SVM. You can use Frequentist methods or Bayesian methods. Or both. It's your choice. Your task is to produce the most useful model you can, one that will help XYZ get the greatest returns from its next campaign.

About the Data:

There are three R binary data files available:

XYZ_customer_data.RData
XYZ_item_data.RData
XYZ_mail_data.RData

A simple R program for reading in the data is provided in the R source code file:

XYZ_read_data.r

And, as you know, and as your luck would have it, an analytic data file has been created for your use:

XYZ_complete_customer_data_frame.RData

The code used to create this file is in the following file. *Be sure to review it and the video describing its construction so that you understand the variables in it:*

miller_xyz_data_integration.r

Data dictionaries are provided in an Excel file (XYZ_Data Dictionaries.xls) and also an Adobe Acrobat pdf file (XYZ_Data Dictionaries.pdf).

Here's what should be a useful tip: when modeling responses, make sure you are using data from customers who were actually targeted by a campaign. Here's another: don't fall into the trap of predicting customer responses using a variable that's another way of measuring the same response.

CART and Other Kinds of Models

Tree models like CART can be used for this kind of predictive targeting exercise. They can also be used to help in identifying predictors that might be useful in other models, sort of like how they could have been used in Solo 1 to ferret out important segment profiling variables when the number of profiling variables is very large.

A main objective when fitting a tree model is to end up with one that performs well in a held out sample of data because overfitting is minimal. This is true of pretty much any kind of predictive model, of course, including the regression models that can be used for this

assignment. Reasonable practice at minimum consists of randomly dividing your sample into estimation and test subsamples (say using a 70/30 split), using your estimation subsample to fit your model that you then evaluate on the test subsample.

Classification and Regression Tree (CART) models are very widely used, and are implemented in the R Core as the rpart library (it's heavily documented), in the tree library(ditto), and in other places in R. See the “machine learning” task view, where you'll see different ways of fitting basic tree models as well as more powerful ensemble methods (e.g. RandomForest, Bayesian model averaging) and other methods.

“mathematicalmonk” on YouTube has some very good introductory videos about CART models. See, e.g.,

https://www.youtube.com/watch?v=zvUOpbgtW3c

Or, search for “mathematicalmonk CART Youtube”

There are Bayesian versions of tree models, e.g. see the R package BayesTree by Chipman et al.

See the chapter about classification and regression trees, and the rpart documentation, in the Solo 3 Huddle.

Finite mixture models, AKA latent class regression models or clusterwise regression models, are also pretty commonly used in applied marketing science applications, and they can be used for response modeling . A way to think about them is that they are regression models in which different subsamples, or clusters, have different regression coefficients, which accommodates a certain degree of customer heterogeneity. The more common algorithms that fit these models simultaneously estimate cluster-specific coefficients and the probabilities that each case belongs to each cluster. Packages in R that do these models can be found in the “cluster” task view. They include packages like flexmix and mixreg that use some version of what's called the “Expectation/Maximization” (EM) algorithm, a method that was originally applied to estimate missing data values. (It still is, too.)

Bayesian methods can be found in the bayesm R package (e.g. rbprobitGibbs()), in the MCMCpack (e.g. MCMCprobit()), and elsewhere. You might find rbprobitGibbs() interesting to use. It uses Albert and Chib's data augmentation trick that we talked about in a Sync. See the Bayesian task view. Bayesian methods fitting mixture regression models can be found in bayesm and in other R packages.

You might find Jackman's book chapter on limited dependent variables useful to review. A copy is available on Canvas. Chapter 9 in Chapman & Feit's “R for Marketing Research and Analytics” (Springer, 2015), includes some coverage of logistic regression.

Identifying Customers to Target and Key Variables

This exercise is about scoring XYZ's customers in terms of the likelihood that they will provide positive net revenue (estimated profit on sales less mailing cost greater than zero) if sent a catalogue, and using the results to decide which customers should be mailed to. To do this you need to decide which customers to score, and how to estimate expected net revenue that will result from mailing to them.

There are several ways that the modeling to do this scoring could be done. For this assignment it'll suffice to first predict the binary responses (bought vs. didn't buy) to *any* mailing of the16th campaign (for the purposes of this assignment, this will simplify your modeling task), and then to use your predictions and the information available to you to estimate net revenue given a catalogue is sent to a customer in XYZ's next campaign.

You know that XYZ makes a 10% profit on items it sells through its catalogues (this doesn't include catalogue cost), and that it costs it $1 to get a catalogue to a customer. You have available historical data for estimating catalogue sales revenues.

The kind of score you want to estimate for each customer who might be mailed is a response-probability adjusted expect net revenue. That is, you want something like this for "mailable" customers, k=1...K:

score for customer "k " = *E*(profit from a catalogue purchase by k) times **p**(purchase by k| k is sent a catalogue) minus cost of sending a catalogue,

where **p** is predicted probability, and *E* means "expected value." Your modeling should provide you with a predicted p for each customer who might be mailed a catalog. You need to decide how best to get a useful estimate of net revenue given the information available to you, an estimate you can rationalize and defend.

Note that in XYZ's mail campaigns, some customers haven't been mailed anything at all, and some have been mailed once or more. We don't know how the decisions have been made regarding how often a customer was mailed to. You're going to treat any mailings as if it's one mailing, to make your modeling task easier.

IMPORTANT: For modeling purposes, you want to use customers for whom the R data variable ANY_MAIL_16 is coded 1.0. These are customers who were sent one or more catalogues. Your response variable will the RESPONSE16.

A deliverable that XYZ would want to receive from you in addition to your report is a list of the account numbers of the customers your predictions indicate XYZ send a catalogue to, i.e, the customers who are likely to produce positive net revenue if mailed to. A traditional direct marketing way of conceptualizing the results is as a list of customer identifiers associated with scores that are used to decide whether to target them or not. When this list is sorted in descending order of the scores, and larger scores reflect greater potential revenue, then the decision of which customers to target is sometimes expressed as "how deep down the list" to mail or to target. **Be sure to include in your report the number of customers to be mailed to who were mailed during the 16th campaign, and also the number of**

customers who were not mailed to in the 16th campaign that your analysis suggests should be mailed to in XYZ's next campaign.

Lastly,...

Last but not least, be sure to follow the formatting and submission instructions provided on Blackboard for this assignment. Post your questions to the Micro 3 / Solo 3 Huddle.

Now it's up to you to see that XYZ Goes Forth and Prospers.