

# Solo 3: Developing an Email Targeting Model

Christina Macholan | Predict 450, Section 55

## STUDY DESIGN

In the following analysis, I explored several classification and regression modeling techniques to determine which customers the XYZ retail company should target in their upcoming mail campaign in order to maximize its cost effectiveness.

XYZ has a database of 30,799 customer records which includes their own data about customers' historical purchase history and response to previous campaigns. They have also purchased Experian data that has been matched up to the individual customer records to expand the available set of predictors based on demographics. In total, the database contains over 500 variables.

To build the classification model to predict the most likely campaign responders, I used the subset of 14,922 customers who received the most recent mailing campaign (campaign 16). Of those that received campaign 16, only about 10% (1,440) responded with a purchase.



## METHODOLOGY & RESULTS

### Data Preparation

With such a large number of variables to sift through, significant attention was paid to data exploration, data clean-up, and variable selection before building any models. In order to reduce the complexity of the data, the following steps were taken:

- Variables that contained information about the response to campaign 16 were removed from the set of predictors since this is data we couldn't have known before predicting the outcome of the response to campaign 16. These include variables like QTY16, AMT16, and BUYER\_STATUS (which essentially shows who has purchased in the last 12 months). Because one of the goals of this modeling process is to be able to predict outcomes for new customers, relying too heavily on very recent response rates to mailers will make the model "short-sighted" and not useful for predicting outcomes when the customer has never received a mailer before.
- Most binary variables were turned into to flag variables with values of 0 or 1.
- Any values that were identified as "unknown" according to the Experian data dictionary (e.g. "", "U", "U0", "00", etc.) were set to NA. After adjusting these records, any variable that had only one factor level remaining were turned into flag variables (1 for "Yes" and 0 for "No").
- Any character variables with more than 50 unique values were removed from the dataset since the algorithms I used do not accept massively categorical variables.
- Any variables with more than 3,000 missing values in the full dataset were removed from consideration.
- Before using algorithms to build the model, any records with missing values were omitted or imputed to the median for numeric variables so that the algorithms could handle predictions on records where missing values were present.

## Data Exploration & Variable Section

Because XYZ is an appliance and electronics retailer, I was especially interested in exploring variables in the data set related to homeownership and interest in electronics. I initially started with a visual exploration of all of the potential predictor variables against the binary variable for responses to campaign 16. During this process, several of the numerical variables were log-transformed to normalize their distributions.

A full set of all the plots created for EDA is included in Appendix A, however some of the most notable variables are shown to the right.

First of all, variables related to a customer's purchase history and previous campaign response are the most highly differentiated between campaign 16 responders and non-responders. Cumulative variables that measure the quantity and amount of purchases through 2009 and variables for the most recent campaign 15 will certainly be important to include in the final selected model.

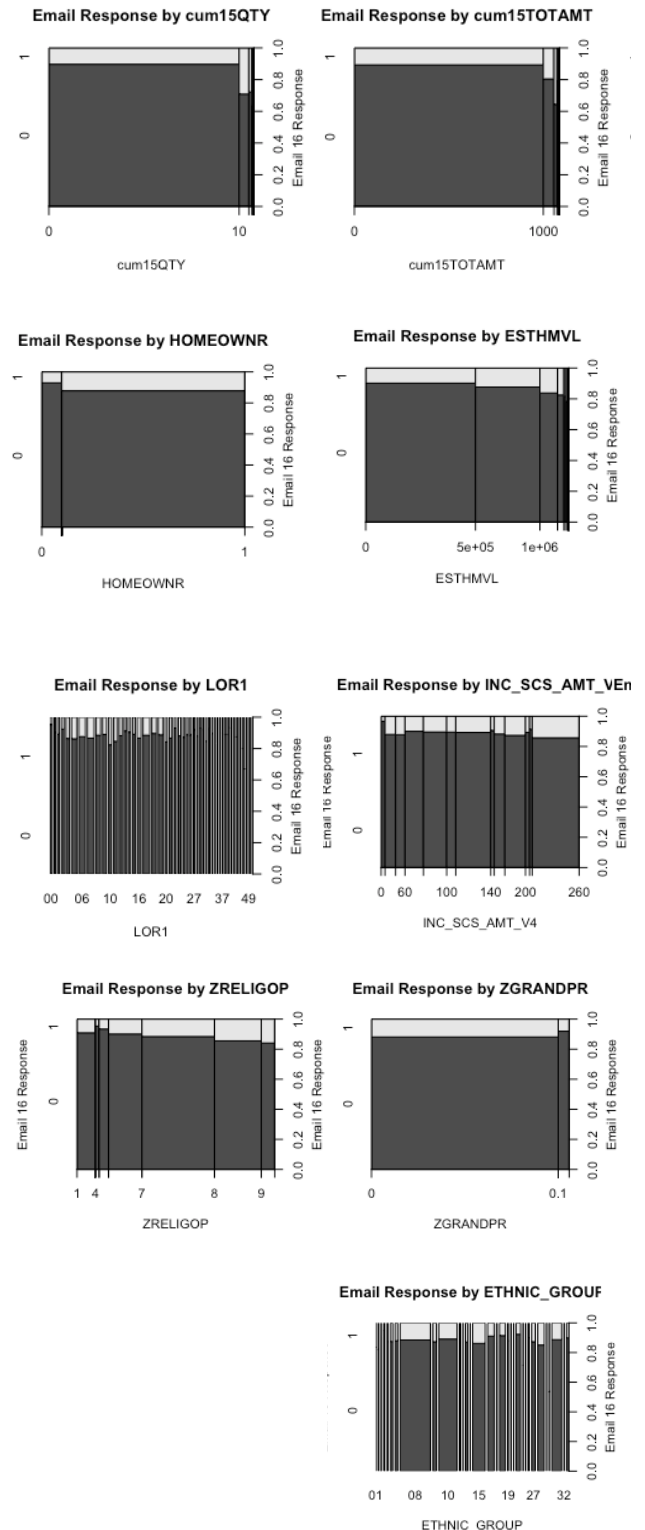
In addition, the second row of plots to the right show a higher likelihood of response for homeowners and customers with higher estimated home values.

The LOR1 plot showed some trends at the extremes -- a higher likelihood to purchase for people who have been in their residence for less than 5 years and a lower likelihood to purchase for people who have been in their residence for 40+ years. I used this information to create two new flag variables for the longest and shortest residence durations.

The data showed trends related to income through many different variables with the high-level insight that customers with higher incomes tend to be more likely to make a purchase and customers below \$20K in income (based on the Experian "probable" demographic data) show a much lower likelihood to make a purchase.

Finally, several demographic and interest variables like ethnic group, language and religion showed some variation between the campaign responders and non-responders that could potentially be predictive. The list that follows shows which variables were retained for the first iteration of modeling. It was important to keep in mind that several of these variables will contain overlapping information (e.g. age variables and the "Grandparent" flag) that are spread across both categorical and numeric variables.

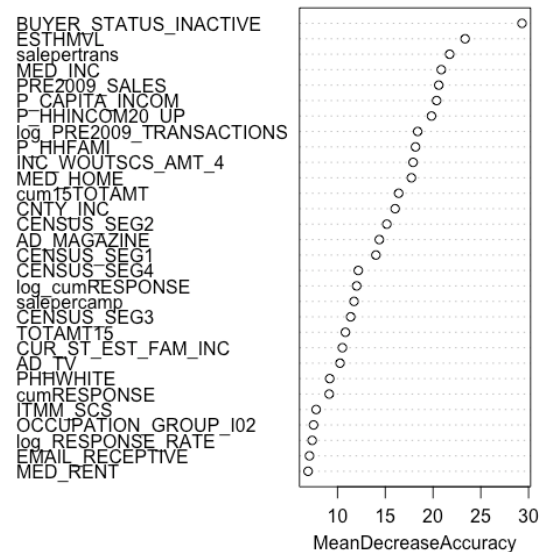
To ease variable selection from the hundreds of options (many of which contain redundant information), I ran algorithms using the Lasso method, and best subset selection to explore the relative importance of each variable in the model. Outputs from these methods were used to select the variables that I then input into the other algorithmic models for predicting the responses to the new mailer.



## Final Variables included in the Models

I attempted to create dozens of new variables in attempt to improve my prediction performance, but the following shortened list of variables proved to perform the best out of all the models and variable combinations explored. They are ordered in the variable importance plot on the right according to the importance in the 100 tree Random Forest model that gave the best predictions on out-of-sample data. Some of the variables appear to be redundant in the model (especially ones related to income), however they were retained because the out-of-sample prediction was good enough to justify better prediction over ease of model explanation.

Variable Importance Plot for best performing Random Forest Model



- **BUYER\_STATUS\_INACTIVE:** A flag variable showing if the user is long-lapsed. (This replaces the three-level factor called "BUYER\_STATUS" that contained an "ACTIVE" status that would have been updated downstream by a response to campaign 16 and would result in an artificially overfit model.)
- **ESTHML:** The estimated value of the customer's home, if they have one.
- **salepertrans:** The average revenue per transaction for the customer. This number will also be used to calculate the average expected revenue from mailer 17.
- **PRE2009\_SALES:** The amount of sales for a customer before 2009.
- **log\_PRE2009\_TRANSACTIONS:** The number of transactions before 2009.
- **cum15TOTAMT:** The total amount the customer has spent up to mailer 15.
- **CENSUS\_SEG:** Segment categorizations of for the household based on the census.
- **cumRESPONSES:** A variable that summed the number of previous responses (pre-campaign 16).
- **ZCOMPUTR:** Customers who showed interest in computers tended to respond more often.
- **EMAIL\_RECEPTIVE:** This flag variable was retained because people who are email receptive tended to be more likely to respond.
- **PHHWHITE:** The percentage of the population in the customer's zip code that identifies as white.
- **ETHNIC\_GROUP:** A few specific ethnic groups were extracted as flag variables.
- **OCCUPATION\_GROUP:** A select set of occupational groups were extracted as flag variables
- **OVER\_55:** A flag variable for customers over age 55. These customers tend to respond less.
- **ADULT\_19\_24:** A flag variable for adult customers under the age of 24. These customers tend to respond more.
- **MCD\_CCD:** A few of these minor civil division codes were extracted as flag variables since they seemed to have predictive power for some neighborhoods that responded far more than others.
- **ZIP:** Similarly, a few zip codes were extracted since they responded more frequently to mailings. I expect that these variables will actually capture latent variables that are not expressed in the Experian data.
- **Several income variables, are included in the model, emphasizing that wealthier families are more likely to respond to the mailer:**

- **MED\_INC:** The median income for the zip code where the customer lives.
- **P\_CAPITA\_INCOM:** Per capita income for the zip code where the customer lives.
- **P\_HHINC20\_UP:** The percentage of the population in the customer's zip code that have an income of \$20K or more. Customers in areas with a lower percentage about \$20 are less likely to respond.
- **CNTY\_INC:** A variable to capture the county-level income of the zip code where the customer lives.
- **INC\_WOUTSCS\_AMT\_4**

## Approach to Evaluating Models

Using a sample of 9,714 training observations, I built a set of classification models with the goal of maximizing expected net profit from the direct mailing campaign. I used the assumption that each mailing costs mailing costs \$1.00 per item and that the average profit from the mailing is 10% of the average transaction value for the audience that will be targeted. The models were each created using a training data set and validated against a set of 5,208 out-of-sample validation observations before selecting the final model.

As a baseline, I set the values for the hypothetical scenario in which all 14,922 customers who received mailing 16 would also receive email 17. Assuming the average revenue per customer is the 1% trimmed mean revenue per customer for campaign 16, we get a value of \$23.45. In this scenario, we would expect \$334,999 in profit from targeting this subset of users if *no model* were used. This is the performance we aim to beat.

In addition to evaluating the models by their out-of-sample performance and potential to predict responses on a validation set, I also looked at model statistics, like relative AIC for models created using the same algorithm, and AUC across all models.

## Model Development

Before beginning the modeling process, I used the Lasso algorithm, Random Forests, and the best subset selection algorithm in R to learn more about which variables carried the most importance for prediction. Using these two techniques, I was able to select the above variables as the best reduced set of variables for building my prediction models.

### Simple Classification Tree

The first model I developed was a simple classification tree, which ended up nearly re-creating the baseline model. The tree had an out-of-sample prediction accuracy of 88.00% and an AUC of 63.03%. It estimated a potential \$38,603 uplift in profit if used versus the baseline model.

### GLM Logistic Regression Models

After selecting the above candidate variables, I first used logistic regression to create a set of models including various combinations of the predictor variables. Models with relatively low expected net profit on the validation data set were discarded. The best performing logistic model was created after reducing the initial 71 variables to a selection of 20 variables chosen by the best subset selection method. The resulting model had an AUC of only 63.03% and an out-of-sample prediction accuracy of 88.00%. Despite this being the best logistic regression model I was able to build, it still did not outperform the baseline model, as shown in the Model Comparison section below.

### Random Forest & Boosted Models

The next set of models were created using the aggregated decision tree methods of Boosting, and Random Forests.

71 variables were input into the Random Forest model, which achieved \$53,875 estimated uplift in profit versus the baseline. The model had only a 50.08% AUC when measured in the out-of-sample data, but a 90.07% predictive accuracy. I attempted to use boosted

decision trees, adjusting the learning rate to try to improve the performance, and this model had worse performance continued to get very low prediction counts without enough lift in the average predicted revenue to make the model worth deploying.

More details about each of the “best” performing models’ performance are summarized below on the following page.

### Model Comparison

The following table shows a comparison of how each of the tested models performed against the scenario where the mailings are sent to all customers.

### Financial Analysis based on All Recipients of Mailing 16

	Classification Tree	GLM 71 Variables	GLM Variables from Best Sub. selection	Random Forest 71 Variables	Boosted Random Forest
Predicted Probability (percentage)	N/A	10% cutoff	5% cutoff	N/A	4% cutoff
<b>XYZ Current Targeting Methods</b>					
Sample Size (All Customers Get Direct Mailing)	14,922	14,922	14,922	14,922	14,922
Average Revenue per Customer	\$23.45	\$23.45	\$23.45	\$23.45	\$23.45
Direct mail cost per Customer	\$1.00	\$1.00	\$1.00	\$1.00	\$1.00
Ave. Revenue Minus Mail Cost per Customer	\$22.45	\$22.45	\$22.45	\$22.45	\$22.45
Profit with of Current Targeting Methods	\$334,999	\$334,999	\$334,999	\$334,999	\$334,999
<b>XYZ Targeting with New Model</b>					
Targeted Customers					
Number of Customers Targeted	14,455	6,282	10,162	14,024	11,208
Average Revenue per Customer	\$25.47	\$28.09	\$26.02	\$25.34	\$24.51
Direct mail cost per Customer	\$1.00	\$1.00	\$1.00	\$1.00	\$1.00
Ave. Revenue Minus Mail Cost per Customer	\$24.47	\$25.75	\$25.75	\$25.75	\$23.51
Profit with New Model	\$353,714	\$161,762	\$261,672	\$361,118	\$263,500
Profit Increase or Loss with New Model	\$18,715	(\$173,237)	(\$73,327)	\$26,119	(\$71,499)
Per Customer Profit Contribution or Loss	\$1.25	(\$11.61)	(\$4.91)	\$1.75	(\$4.79)
Number of Customers in Database	30,779	30,779	30,779	30,779	30,779
Estimated Profit Contribution/Lift of Targeting	\$38,603	(\$357,330)	(\$151,249)	\$53,875	(\$147,478)
<b>Out-of-Sample Accuracy</b>					
AUC	88.00%	63.02%	62.62%	90.07%	88.69%
	63.03%	75.77%	76.38%	50.08%	42.17%

## APPLYING THE FINAL MODEL

### Predicting Outcomes for Campaign 16 Recipients

In order to apply this final model to XYZ's full database, a few additional steps need to be taken:

- **Replace campaign 15 variables with campaign 16 variables.** Several of the variables used in the prediction model take into account the most recent campaign's performance (campaign 15). However, in order to predict campaign 17's performance, it makes sense to use the values from campaign 16 in place of the campaign 15 variables that were used to build the model.
- **Add campaign 16 values to cumulative variables.** Several of the variables used in the prediction model take into account the full purchase history and campaign performance history for the customer. In order to predict outcomes for campaign 17, campaign 16 values should be added to the cumulative prediction variables first.

### Predicting Outcomes for Other Customers

- **Impute missing values in the customer database.** Some customer records are missing values for the variables that were used in the final selected model. In order to deploy the model across all customer records, we need to come up with a set of decision rules on how to treat any missing values. These values can be imputed based on the customer's similarity to other customers, or the customers could be treated with a different, simpler model.

In some cases, if there are customers who have too much missing information to deploy the model directly on their records, an alternative would be to produce a KNN model using a limited, known set of variables in order to have a "best guess" of which of these customers may be the most likely to respond. We could also develop some sort of typing tool (similar to what we discussed in our segmentation project) to align our targeting for those users with the overall targeting model.

### Model Improvements & Limitations

In this project I only scratched the surface of the many algorithms that could be used to make classification predictions. Other models including Linear Discriminant Analysis, Quadratic Discriminant Analysis, SVMs, and Bagging could have perhaps performed better than the models explored in this iteration.

Additionally, it was new for me to learn this week that there are algorithms especially designed to handle data that is aggregated at different levels (e.g. individual level and zip code level) like the data set we used for this modeling exercise. Because zip codes did seem to have some predictive capabilities in the models I developed, this would be worth exploring further to ensure the models built are stable and align with the assumptions of the algorithms that are used.

# APPENDIX

A.

