Marketing Campaign Response Modeling

Northwestern University
Andrew Kang
MSDS 450

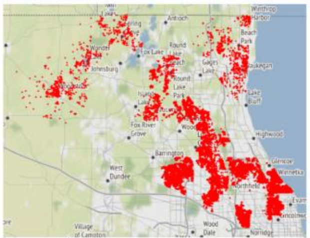
Introduction

XYZ is looking to develop a model-based methodology to target customers with mailers by combining customer data with demographic and geographical data. The dataset that was used for this analysis contained 30,779 observations and 554 variables. Given the high dimensionality of the dataset and the behavioral focus of the use case, careful consideration was required during exploratory data analysis and data pre-processing.

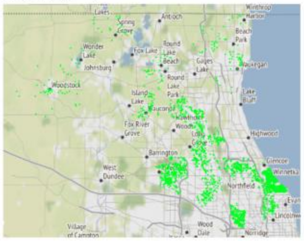
Exploratory Data Analysis

In order to study the effects of advertising on purchase propensity of respondents, the dataset was split based on whether a customer received a mailer in the last campaign. Then, observational data was analyzed and compared against various features. Based on a geographic analysis of responses, we can see that customers that responded to the mailer tend to be clustered towards the urban areas of Chicago. This likely points to geographical variables being an important factor, and engineering features based on distance from the urban area is worth exploring further in future analysis.

Response16 = 0



Response16 = 1



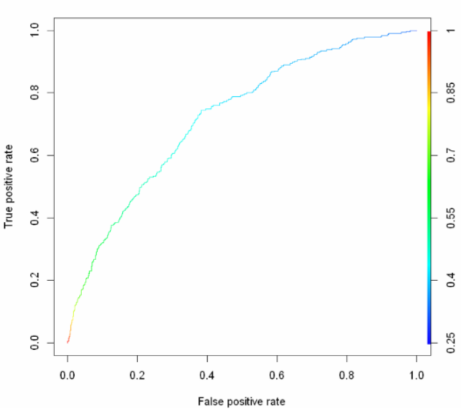
In order to further improve the models, domain expertise was also used to helped guide the variable reduction process. In the end, the selected variables for our models fit into three key categories: income, historical sales, and family. These categories were used to help bridge the gap in terms of logical relationships between variables and the response variables within the dataset. In the end we selected the below variables to begin the modeling process.

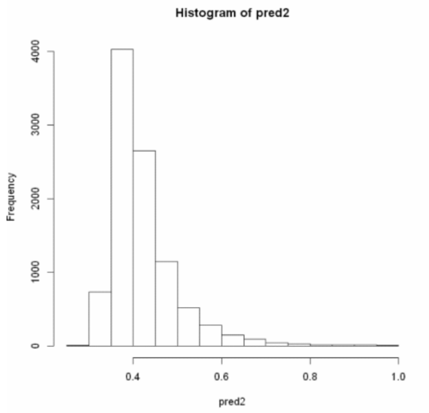
Variable	**Description**
'cum15QTY'	Quantity Through Campaign 15
'cum15TOTAMT'	Total Amount Through Campaign 15
'EXAGE'	Exact Age
'HOMEOWNR'	Homeowner
'MED_INC'	Median Income
'PRE2009SALES'	Sales Before 2009
'RESPONSE16'	Response Variable
'salepercamp'	Sales per Campaign
'salepertrans'	Sales per Transaction
'TOTAMT15'	Campaign 15 Total Amount
'PRE2009TRANSACTIONS'	Transactions Before 2009
'ZCREDIT'	Presence of Credit Card
'NAT_INC'	National Income Percentile
'LOR1'	Length of Residence
'NUMBADLT'	Number of Adults
'NUM_CHILD'	Number of Children
'ZONLINE'	Presence of Internet

In terms of data cleansing, the variables came in a variety of formats and we had to address some formatting and logical challenges with the way that the data was organized. From the initial 33,079 observations, sales values for customers with no purchase history were set to 0 to avoid infinite values and nulls. One reason this was important was because we did not want to throw away values for customers that did not make a purchase. In fact, those values would prove to be helpful in testing our model against cases where mailers have led to no historical sales or transactions. There were several cases where variables that were modeled were rejected in favor of raw data that could be more objectively analyzed. Once we completed our variable selection, factorization, and data cleansing process, we were left with 9,784 observations with a non-response rate for Campaign 16 of 88.97%. Given the heavy imbalance towards non-response, one of the key steps we took was to overbalance the training data using the ROSE R package. By oversampling, we were able to keep the full set of non-response data while forcing the model to make actual predictions above a random guess.

Logistic Regression

As the predictions required classification of response or non-response, we start with a simple model to help understand the nuances and challenges that could be addressed with more nuanced models. Initially, we fed in all of our variables into the Logistic Regression model for an initial fit. One of the key insights here was that our ZONLINE, ZCREDIT, and HOMEOWNR variables were all showing up as the same value in our dataset. Consequently, we can assume that these factors were likely already being targeted in some way by the campaign. As a result, we removed these variables and re-fit the logistic regression model. With our test holdout of 30%, we had an accuracy of ~76% and an AUC of 68.55%. The ROC curve shows that we can create results that are better than random, but the model still struggles with precision and recall. The histogram shows how our results are skewed toward 0 values, which is to be expected.

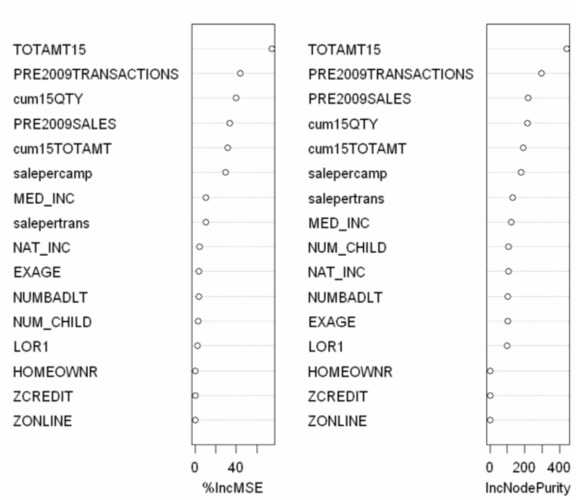




While the performance was quite good out-of-sample given that the data was balanced, there is still less than optimal performance in terms of predicting actual customers that received the mailer. This could be due to limitations with the regression modeling approach, which is why we decided to also model a decision tree approach using random forests.

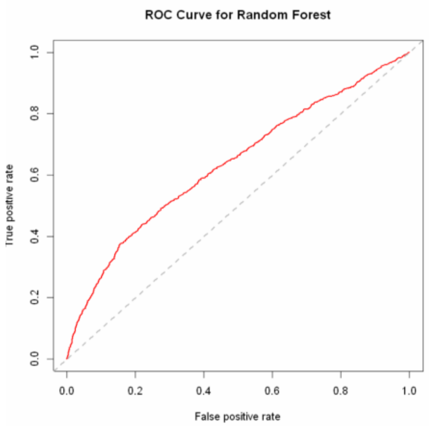
Random Forest

With our Random Forest model, we utilized the same training data set with balanced data. In the model, we were able to extract variable importance. Based on the plots below, we reached the same conclusion as with our logistic regression example. Inevitably, while HOMEOWNR, ZCREDIT, and ZONLINE may indeed be predictive, the targeting of the 16th campaign means that we cannot extract any predictive value from our analysis.



With our model, we can see that the previous sales and transaction information is consistently appearing at the top in terms of variable importance. The fact that TOTAMT15 shows up at the top suggests that the consumers that have been recently active in responding to the mailer are again responding to the 16th mailer.

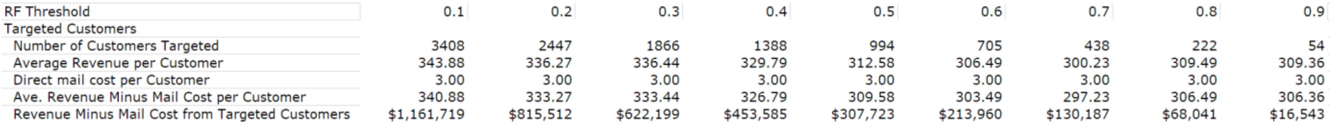
The ROC curve, accuracy score, and AUC score of our reduced model after removing HOMEOWNR, ZCREDIT, and ZONLINE shows an improvement over our logistic regression model.



The AUC score of the RF model was 69.45% and our accuracy score jumped to 85%. Based on the improved results, we chose to continue with the Random Forest version of the model as our selection.

Targeted Customers

Given that we have selected our random forest model, we can now make financial projections for targeted advertising. Below, we have listed the expected net profit for each of the RF thresholds. One thing to note is that it appears that a lower threshold yields more profitability overall. This is a key insight as our accuracy score is less important than our ability to correctly balance customers and high value purchasers.

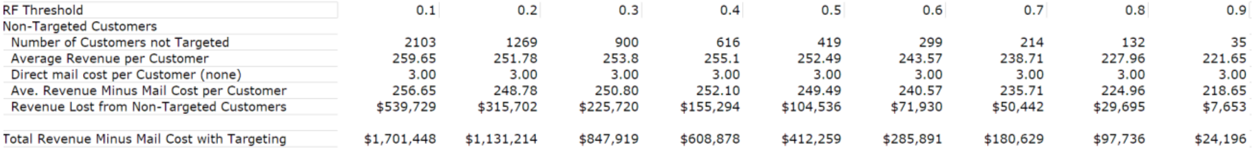


RF Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Targeted Customers									
Number of Customers Targeted	3408	2447	1866	1388	994	705	438	222	54
Average Revenue per Customer	343.88	336.27	336.44	329.79	312.58	306.49	300.23	309.49	309.36
Direct mail cost per Customer	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	340.88	333.27	333.44	326.79	309.58	303.49	297.23	306.49	306.36
Revenue Minus Mail Cost from Targeted Customers	$1,161,719	$815,512	$622,199	$453,585	$307,723	$213,960	$130,187	$68,041	$16,543

Non-Targeted Customers

For customers that were not targeted in the last mailer, we scored our model against those that did not receive any mail from the 16th campaign. The objective is to see if we can extrapolate our insights from the targeted customers across the universe of customers and get value. Unfortunately, this approach comes with several limitations. First and foremost, we cannot directly observe the effect of the mailer on purchases. If the purchases were to happen from those that were not targeted anyway, we may see a case where the additional cost of the mailers would not justify an increase in revenue. Additionally, it is important to note that our training set forced us to remove several variables during the training process. Those decisions may or may not have been made if the model was trained across the broader population.

With that being said, the results from the RF model on the non-targeted customers shows lower average revenue across the different RF thresholds. Even so, the model shows positive net profit at each of the thresholds.

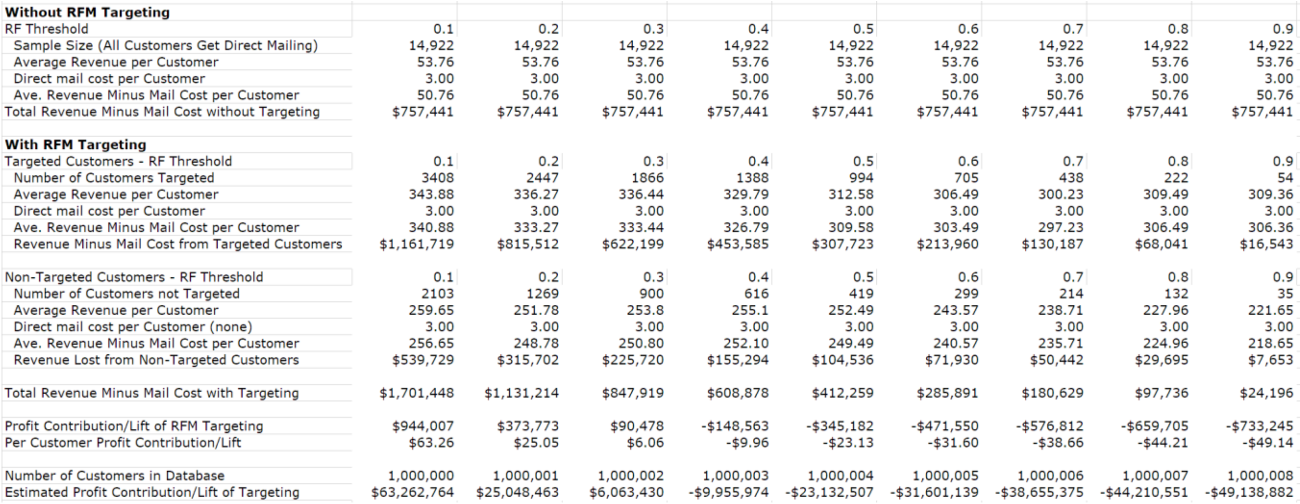


RF Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Non-Targeted Customers									
Number of Customers not Targeted	2103	1269	900	616	419	299	214	132	35
Average Revenue per Customer	259.65	251.78	253.8	255.1	252.49	243.57	238.71	227.96	221.65
Direct mail cost per Customer (none)	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	256.65	248.78	250.80	252.10	249.49	240.57	235.71	224.96	218.65
Revenue Lost from Non-Targeted Customers	$539,729	$315,702	$225,720	$155,294	$104,536	$71,930	$50,442	$29,695	$7,653
Total Revenue Minus Mail Cost with Targeting	$1,701,448	$1,131,214	$847,919	$608,878	$412,259	$285,891	$180,629	$97,736	$24,196

One positive sign that the model can extrapolate is the fact that the number of customers targeted is less than with our targeted customers. The lower number of customers and the lower amount of revenue per customer indicates that there is some acknowledgement that these customers are less likely to respond to new advertising.

Financial Analysis

Now that the model has been applied to both targeted and non-targeted customers, we can now compare net revenue and explain any differences that we observe.



Without RFM Targeting									
RF Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Sample Size (All Customers Get Direct Mailing)	14,922	14,922	14,922	14,922	14,922	14,922	14,922	14,922	14,922
Average Revenue per Customer	53.76	53.76	53.76	53.76	53.76	53.76	53.76	53.76	53.76
Direct mail cost per Customer	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	50.76	50.76	50.76	50.76	50.76	50.76	50.76	50.76	50.76
Total Revenue Minus Mail Cost without Targeting	$757,441	$757,441	$757,441	$757,441	$757,441	$757,441	$757,441	$757,441	$757,441
With RFM Targeting									
Targeted Customers - RF Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Number of Customers Targeted	3408	2447	1866	1388	994	705	438	222	54
Average Revenue per Customer	343.88	336.27	336.44	329.79	312.58	306.49	300.23	309.49	309.36
Direct mail cost per Customer	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	340.88	333.27	333.44	326.79	309.58	303.49	297.23	306.49	306.36
Revenue Minus Mail Cost from Targeted Customers	$1,161,719	$815,512	$622,199	$453,585	$307,723	$213,960	$130,187	$68,041	$16,543
Non-Targeted Customers - RF Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Number of Customers not Targeted	2103	1269	900	616	419	299	214	132	35
Average Revenue per Customer	259.65	251.78	253.8	255.1	252.49	243.57	238.71	227.96	221.65
Direct mail cost per Customer (none)	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	256.65	248.78	250.80	252.10	249.49	240.57	235.71	224.96	218.65
Revenue Lost from Non-Targeted Customers	$539,729	$315,702	$225,720	$155,294	$104,536	$71,930	$50,442	$29,695	$7,653
Total Revenue Minus Mail Cost with Targeting	$1,701,448	$1,131,214	$847,919	$608,878	$412,259	$285,891	$180,629	$97,736	$24,196
Profit Contribution/Lift of RFM Targeting	$944,007	$373,773	$90,478	-$148,563	-$345,182	-$471,550	-$576,812	-$659,705	-$733,245
Per Customer Profit Contribution/Lift	$63.26	$25.05	$6.06	-$9.96	-$23.13	-$31.60	-$38.66	-$44.21	-$49.14
Number of Customers in Database	1,000,000	1,000,001	1,000,002	1,000,003	1,000,004	1,000,005	1,000,006	1,000,007	1,000,008
Estimated Profit Contribution/Lift of Targeting	$63,262,764	$25,048,463	$6,063,430	-$9,955,974	-$23,132,507	-$31,601,139	-$38,655,375	-$44,210,551	-$49,138,882

Overall, we find that our model can improve on sending mailers to all customers up to a certain threshold. While our model does relatively weak in terms of precision, the classification is enough to help improve on no model at all. In fact, the model shows that the best results come from having a low random forest threshold where we are sending mailers to more targeted customers and more non-targeted customers. This likely indicates that there are customers that are extremely unlikely to respond to the mailer, which can be removed with a low threshold.

One key factor to consider in addition to this analysis is the variability in profitability. Based on information available at the time of analysis, the profit for XYZ was 10% on items through the catalogue. If this profit margin remains static, then the process of selecting a threshold is straightforward. However, if it varies according to inventory levels and seasonality, there may be a risk of running an unprofitable campaign in the future.

Adapting the Model

In terms of adapting this model for new customers, one strategy could be to build a propensity model based on the characteristics of existing customers with similar demographic and geographic traits. What we noticed in our initial EDA was that there was a geographic connection between the responses for campaign 16. We ended up throwing away observations that had null values, but we could go back and impute them more confidently as we have a better foundation for analysis.

As we will now have a 17th campaign, there are two key things that we would like to identify through A/B testing. For the targeted customers, we would like to see what happens to a group of customers when we remove the mailer in the next campaign. For the non-targeted customers, we would like to see what happens to a group of customers when we add a mailer in the next campaign. To ensure that we can have fair baselines for comparison, clustering and nearest neighbor approaches should be used to ensure that there are good controls for our test. What we would be looking for is a deviation between our control and test respondents in each of these cases.

Conclusion

While one of our objectives was to produce the best model, the key insight here was that the metric that was most important was net revenue. This was accomplished by slightly reducing the number of mailers based on our Random Forest model. If we focused on precision, we may have likely missed out on a large pool of customers that would have responded profitably. If we focused on recall, we would have reduced our profit by sending out too many mailers.

For XYZ, the appropriate next step would be A/B testing along with a propensity model for new customers. With the combination of several models, XYZ should be able to target consumers both effectively and profitably by combining customer information with geographic and demographic information.