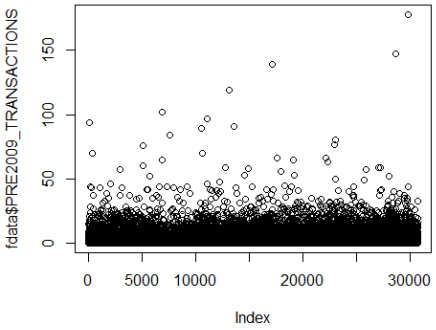
Derek Hughes
PREDICT 450 – Sec 55
Solo #3
Summer 2015

INTRO

Improving target and direct mail marketing can be a challenge for any business. XYZ, a well-known retail company, is interested in improving its ability to identify potential buyers to increase market penetration with its marketing activities. Specifically, XYZ wants to develop a method to target customers in future mail campaigns. To improve their understanding of the market, XYZ has accumulated customer data, recent campaign data, and transaction data. We will use this quantitative research data to design a model for identifying potential buyers, a model for improving (hopefully) the direct marketing campaign that estimates the individual's likelihood of buying, and the estimated financial impacts from using a new model compared to the currently used model.

Data

The raw customer data, recent direct marketing campaign data, and transaction data are separate. We combined the three into one data set that contains approximately 30,779 observations and includes 554 variables. After removing responses containing missing values and those observations that had more than 32 transactions before 2009, we were left with 26,172 observations.



The data set has particularly large dimensions (554) and some information about the data set is either inaccurate or missing, making the task of identifying proper variables to use as response and basis variables more difficult than usual. Furthermore, the data contains both numeric and character variable values and the character variables will need to be converted to dummy variables for analysis and model participation. Extensive exploratory data analysis was required for this type of data set.

To reduce the dimensionality of the data set, we first selected a response or target variable that could be used for both identifying potential buyers from non-buyers and as a buyer for the direct mailing campaign. In this particular case, we created a cust_purchase variable that used "1" to identify a buyer and "2" for non-buyers based off whether or not the customer_status variable was "active" or "inactive/lapsed." The selection of the response variable was determined by logical conclusions based on the assumed overall goal of the targeting program to segment buyers from non-buyers.

```
qty_flg                totamt_flg             cust_purchase
    1     2                1     2                1     2
13602 17177            13602 17177            13602 17177
```

TOTAMT and QTY were also considered as response variables, but they produced identical binomial flag values compared to customer_status so we simply decided to use customer_status for both models.

Next we created a new data base that contained only the response variable and twenty plus basis variables.

These bases variables were determined using a variety of approaches:

Some were selected based on **logical relationships with the response variable**:
PRE2009_TRANSACTIONS are logically related to whether or not a customer is a buyer

Others were a **combination of two related variables**:
MASTERCARD is a combination of MS_PREM and MS_REG variables

Many were selected based on **simple logistic correlations with the response variable**:

```
(Intercept) -2.964e-01  3.720e-02  -7.966 1.63e-15 ***
MED_INC     -3.145e-06  3.623e-07  -8.681  < 2e-16 ***

(Intercept) -0.45453    0.04298 -10.575  < 2e-16 ***
HOMEOWNRY   -0.15936    0.04587  -3.474 0.000513 ***
```

Some were created to be **binary indicators of a specific value in another variable**:
ANY_MAIL_1 is a binary indicator if there is any value > 0 in SUM_MAIL_1

This process was repeated for the model identifying buyers versus non-buyers and the model for direct marketing campaign improvements.

For the buyer segmentation model, we used **these variables as our initial model**:
cust_purchase, PRE2009_TRANSACTIONS, MASTERCARD, MED_INC, HOMEOWNR, RENTER, SUM_MAIL_1, ANY_MAIL_1, EXAGE, ZMUSCRST, ZKITCHEN, NUMBADLT, ZINVESTR, HH_CONTACT_LENSES, ZCRAFTS, ZCOLLECT, ZPRCHPHN, ZTENNIS, ADULT1_G, ZMOBGRDN, SPORTS_RELATED

For the direct marketing campaign model, we used a combination of the aforementioned variables with these additional purchase related variables after testing them against the response variable in single and multiple variable logistic models:

HEALTH_RELATED, ENVIRONMENTAL, ZSPORTS, RE_CURRENT, STOCK_BOND_CURRENT, ZCREDIT, VETERAN, CD_MM_CURRENT, ZONLINE, ZSPENDER, ZMOB

Modeling Procedures

For the first model that is to be used for segmentation of buyers and non-buyers, we used Classification and Regression Trees. These trees use the Gini index to determine

separation points at each node of the tree. The Gini index is a measure of impurity and we are looking for divisions in the tree that create the largest differences between the amount of buyers and amount of non-buyers in the resulting node. The trees should evaluate each variable in the model and, if relevant to improving the impurity at each node based on a predetermined level (maximizes classification/minimizes error), include all variables that improve the impurity. With enough relevant variables, we should end up with a tree with a number of nodes and "branches." In this case, we tested three models, each with different variables, in order to find which produced the most accurate split or segmentation of buyers and non-buyers.

Trees tend to overfit the data; hence we can split the data into training and test data, implement the use of cross-validation, and/or select specific branches that prune the tree to a more reasonable level. We used all three techniques to combat overfitting for each model. When we prune a tree, this creates a subset of data that we can use for predicting the probability that an individual will buy based on the variables we selected to model the direct mail marketing campaign.

The advantage of using a classification tree is that it can be used to characterize outcomes based on many predictors; it can handle high dimensions; it is simple for management or those less knowledgeable about statistics and modeling to understand, but it still retains an ability to produce powerful and accurate results. Finally, tree models are often considered to mimic real life decision making processes more than other models. The main disadvantage of having a tendency to overfit the data has already been addressed, but we should also note that trees are sensitive to small changes to input data, which may produce large changes to the tree structure.

We used a logistic regression model to determine the probability an individual will purchase from the direct mail marketing campaign. The advantage of using the logistic regression was the ability to transform the response variable output (which is in terms of the log of the odds ratio) to a probability. This probability was the likelihood that an individual will buy based on the direct marketing campaign. Furthermore, we used this probability value to determine the financial impact of only sending direct mail to individuals of a particular probability level. By using various probability levels as a "cutoff" value, we calculated the financial impacts of sending direct mail only to individuals at or above those probability levels. This provided us with the ability to evaluate different scenarios in order to find which, if any, worked best in terms of improving profitability over the conventional "send to everyone" approach.

Again, as with the tree model, we evaluated five different models for the logistic regression using the subset we selected from the decision tree model. After selecting the most accurate model of the five, we used three different cutoff points to determine which approach would most likely yield the maximum profitability from a direct mail marketing campaign.

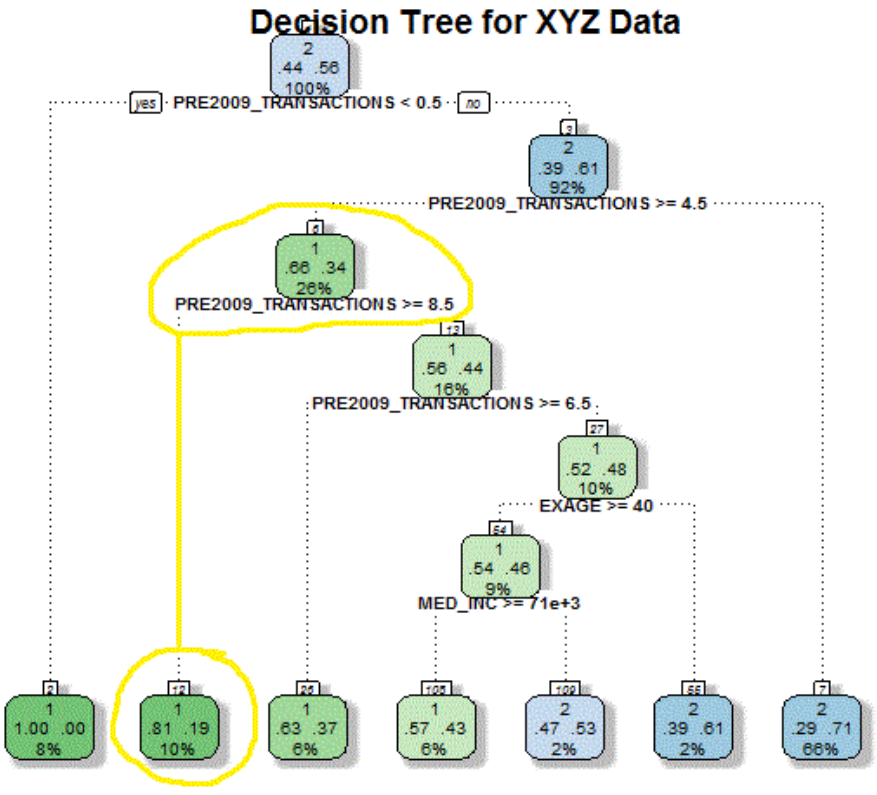
Results

BUYING CUSTOMER SEGMENTATION

Our first classification tree model used to segment buyers from non-buyers included the full gamut of variables we selected. From the visual tree plot we can see that there were 11 of nodes and variables used in the tree. Our terminal node of interest (12) contained 867 buyers and 169 non-buyers for a total of 1036 people. Note we are using the training data set which includes 12,990 people instead of the 30,000+ in the full data set. The percent of people in this node of all customers is 8% percent. And out of that number of people, 84% percent or 867 people bought something.

When we check the area under curve, it shows .7347 AUC. By using a confusion matrix to compute the accuracy of the model, we find it produces a .7212 of accuracy.

However, our main goal is to test the validity of the model. We do this by testing the model using the "holdout" or test data. When we check the area under curve for the test data, it shows .7339 AUC. The confusion matrix for the test data produced .7217 accuracy.



By comparison, we can see that the model performed better on test data. We will use the test data results to compare to the other three models.

Our second classification tree model used to segment buyers from non-buyers included only the variables identified as importance from the first model plus ZONLINE and ZSPENDER. From the visual tree plot we can see that there were 11 of nodes and variables used in the tree. Our terminal node of interest (12, again) contained 867 buyers and 169 non-buyers for a total of 1036 people. The percent of people in this node of all customers is 8% percent. And out of that number of people, 84% percent or 867 number of people bought something.

When we check the area under curve, it shows .7347 AUC. By using a confusion matrix to compute the accuracy of the model, we find it produces a .7212 of accuracy.

When we check the area under curve for the test data, it shows .7339 AUC. The confusion matrix for the test data produces .7217 accuracy.

By comparison, we can see that the model performed better on test data, but we can also see its results were identical to the full model. Hence we can use fewer variables to produce the same results but our additional variables (ZONLINE and ZSPENDER) had zero effect. We will use the test data results to compare to the other three models.

Our third classification tree model to segment buyers was the same as model two but with the addition of the ZMOB variable. The training data results were the same in terms of nodes, best terminal node, accuracy, and AUC. The test data results were the same as well; hence, we will use the second model because it contains the fewest variables which will reduce model complexity.

One item of note is the difference between the training and test set results. Thirteen nodes and variables were used in the test data set tree. Our terminal node of interest contained 1095 buyers and 262 non-buyers for a total of 1357 people. The percent of people in this node of all customers is 10% percent. Here only 81% bought something, as opposed to 84% for the other tree but the total number of buyers increased with this terminal node.

Since both the training and test data set produced very similar results, it lends credence that our model is valid over different data sets and doesn't overfit the data.

Using pruning on the tree, we decided to prune at node five. This is where the condition is based on the PRE20009_TRANSACTIONS variable where the number of transactions per individual is greater than or equal to 8.5. By pruning our selected tree, we can maintain the accuracy, reduce the complexity, and reduce the model's ability to overfit the data. This pruned branch will become our subset data for our mail campaign model.

MAIL CAMPAIGN MODEL

Using the subset data based on the results from our tree evaluation, we created four logistic regression models from which to compare.

Our first logistic regression model used to model the probability that an individual will make a response/purchase from the direct mail campaign included all variables in the full subset data. The train and test AIC for this model was 1204 and 692 respectively.

Test data:

PRE2009_TRANSACTIONS	-1.425e-01	1.821e-02	-7.826	5.04e-15	***
MASTERCARD2	-2.668e-03	1.277e-01	-0.021	0.9833	
MED_INC	-2.226e-06	1.381e-06	-1.613	0.1069	*
HOMEOWNR2	2.752e-01	2.382e-01	1.155	0.2481	
RENTER2	-4.862e-01	5.667e-01	-0.858	0.3910	
SUM_MAIL_1	-8.685e-01	9.125e-01	-0.952	0.3412	
ANY_MAIL_12	-6.885e-01	9.271e-01	-0.743	0.4577	
EXAGE	1.687e-03	4.543e-03	0.371	0.7103	
ZMUSCRST2	2.724e-01	4.397e-01	0.620	0.5356	
ZKITCHEN2	1.520e-01	1.355e-01	1.122	0.2620	
NUMBADLT	3.894e-02	4.337e-02	0.898	0.3693	
ZINVESTR2	1.728e-01	1.459e-01	1.185	0.2361	
HH_CONTACT_LENSES2	-1.134e-01	2.724e-01	-0.416	0.6772	
ZCRAFTS2	-8.005e-02	1.466e-01	-0.546	0.5851	
ZCOLLECT2	-3.818e-01	1.839e-01	-2.076	0.0379	*
ZPRCHPHN2	9.118e-01	1.630e+00	0.559	0.5758	
ZTENNIS2	3.035e-01	4.736e-01	0.641	0.5216	
ADULT1_G2	-4.214e-02	1.597e-01	-0.264	0.7919	
ADULT1_G3	-2.513e-01	2.566e-01	-0.979	0.3275	
ZMOBGRDN2	3.175e-02	1.713e-01	0.185	0.8530	
SPORTS_RELATED2	3.260e-01	1.427e-01	2.284	0.0224	*

For the next logistic model, we included only those variables that were significant at a .11 level in the first model in either the training OR test set. This is not normal practice but considering the small sample size and large differences between which variables were significant based on the test or training set, we decided to include all variables that were significant in either because this model is an intermediate step for building our final model. This included a total of 7 variables. The train and test AIC for this model were 1361 and 776 respectively.

For the third logistic model, we added new variables, which were either significant in a simple logistic regression or were purchase related, to the second model's significant variables. This included: ENVIRONMENTAL+ ZCREDIT+ STOCK_BOND_CURRENT+ RE_CURRENT+ ZSPORTS+ CD_MM_CURRENT+ ZONLINE+ ZSPENDER+ ZMOB
The train and test AIC for this model were 1378 and 792 respectively.

For the final logistic model, we ran a backward variable selection process on the third model and found even more model improvement. The AIC for the model improved to 1361 and 774 for the train and test set, respectively.

The final model was:
cust_purchase ~ MED_INC+HOMEOWNR+ZKITCHEN+ZMOBGRDN+ZCOLLECT+
SPORTS_RELATED

By comparing the AIC values for each model, we selected the final model because, while it's AIC was not as low as using the full model, it used significantly fewer variables (7 vs. 22) which improves parsimony, complexity, and greatly improves interpretability.

FINANCIAL COMPUTATIONS

Based on our selected model, we will now arbitrarily create five different scenarios from which to determine the financial implications of each scenario compared to the current approach used by the XYZ company. We will create the five different scenarios by using different cutoff values based on the predicted probability that the individual will purchase a product.

```
> dim(test_data5)
[1] 2833   33
> sum(pred2>.30)
[1] 416
> sum(pred2>.20)
[1] 1543
> sum(pred2>.15)
[1] 2042
> sum(pred2>.10)
[1] 2408
> sum(pred2>.05)
[1] 2624
```

Our first scenario only uses individuals who had a 30% probability of making a purchase after the most recent direct mail campaign.

The results show that 416 customers would be targeted. The average revenue per customer (minus mail costs) would be $52.40. Hence, the profit for scenario one would be $20,550. Comparing to XYZ's current targeting methods, this represents a $33,673 decrease in profitability.

Our second scenario only uses individuals who had a 20% probability of making a purchase after the most recent direct mail campaign.

The results show that 1,543 customers would be targeted. The average revenue per customer (minus mail costs) would be $47.97. Hence, the profit for scenario two would be $74,018. Comparing to XYZ's current targeting methods, this represents a $19,794 increase in profitability.

Our third scenario only uses individuals who had a 15% probability of making a purchase after the most recent direct mail campaign.

The results show that 2042 customers would be targeted. The average revenue per customer (minus mail costs) would be $63.59. Hence, the profit for scenario three would be $123,725. Comparing to XYZ's current targeting methods, this represents a $69,501 increase in profitability.

Our fourth scenario only uses individuals who had a 10% probability of making a purchase after the most recent direct mail campaign.

The results show that 2408 customers would be targeted. The average revenue per customer (minus mail costs) would be $60.59. Hence, the profit for scenario three would be $145,901. Comparing to XYZ's current targeting methods, this represents a $91,677 increase in profitability.

Our fifth scenario only uses individuals who had a 5% probability of making a purchase after the most recent direct mail campaign.

The results show that 2624 customers would be targeted. The average revenue per customer (minus mail costs) would be $26.59. Hence, the profit for scenario five would be $69,772. Comparing to XYZ's current targeting methods, this represents a $15,549 increase in profitability.

	Logistic Regression Model Cutoff Rules				
Predicted Probability (percentage)	30.00	20.00	15.00	10.00	5.00
XYZ Current Targeting Methods					
Sample Size (All Customers Get Direct Mailing)	2833	2833	2833	2833	2833
Average Revenue per Customer	22.14	22.14	22.14	22.14	22.14
Direct mail cost per Customer	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	19.14	19.14	19.14	19.14	19.14
Profit with of Current Targeting Methods	$54,224	$54,224	$54,224	$54,224	$54,224
XYZ Targeting with New Model					
Targeted Customers					
Number of Customers Targeted	416	1,543	2,042	2,408	2,624
Average Revenue per Customer	52.40	50.97	63.59	63.59	29.59
Direct mail cost per Customer	3.00	3.00	3.00	3.00	3.00
Ave. Revenue Minus Mail Cost per Customer	49.40	47.97	60.59	60.59	26.59
Profit with New Model	$20,550	$74,018	$123,725	$145,901	$69,772
Profit Increase or Loss with New Model	($33,673)	$19,794	$69,501	$91,677	$15,549
Per Customer Profit Contribution or Loss	($11.89)	$6.99	$24.53	$32.36	$5.49
Number of Customers in Database	30,779	30,779	30,779	30,779	30,779
Estimated Profit Contribution/Lift of Targeting	($365,841)	$215,052	$755,092	$996,022	$168,926

After reviewing all five scenarios, we found that the scenario using 10% predicted probability produced the best result because it had the highest profit increase over the old model and highest per customer profit contribution. Compared to the current targeting methods used by XYZ, this scenario is a better model than sending mail to every customer.

CONCLUSION:

The results of this analysis show that the profile of a buyer versus a non-buyer should be targeted mainly by the number of transactions the customer had before 2009, specifically if they were greater than or equal to 4.5 transactions. There were certainly other variables that could be used as well to profile the buyer, such as median income greater than or equal to 71K and age greater than or equal to 40, but our philosophy is the simpler the better so we recommend profiling by simply using pre-2009 transactions of greater than or equal to 4.5 transactions.

In regard to which customers XYZ should target for their direct mail campaign, we recommend a customer who has greater than or equal to 4.5 transactions before 2009, is a homeowner, has kitchen aids/small appliances, is a collector, has made mail order purchases of garden equipment/supplies, and has made sports related purchases in the past.

When using the mail order campaign model, we recommend using it for buyers who have made more than 4.49 transactions before 2009 to improve its accuracy. Furthermore, we recommend only sending mailers to those individuals with a predicted probability of purchasing of 5% to 20% to improve profitability. However, ultimately we specifically recommend only targeting those with a 10% predicted profitability of purchasing. At that level, XYZ company should maximize its profitability over sending mail to every customer in its database.