TO: Professor Anil Chaturvedi

FROM: John Hokkanen

DATE: May 29, 2017

RE: Marketing Project Predictions

EXECUTIVE SUMMARY

The assignment requires the development of models to optimize profits of a mailing. First a classification model must predict which persons should receive a mailing and a second prediction model will estimate how much they are likely to donate. The classification model is more difficult and more important in this context and so much more effort was placed on that modeling task. The modeling effort focused on intense understanding of the relationships between the variables. To understand what variables were important and in which ways they were related, significance in regressions, tree model variable selection, importance criteria in random forest models, and cross tabulations were all used. In the end, final variables and models were selected on the basis of profitability criteria against the validation data. The classification model put forth is an LDA model using a variety of supplied predictors, derivative categorical dummies, and newly created categorical interaction variables.

The donation estimation model selection was easier to develop. After using best subset analysis to generate a variety of variable selections, a baseline MSE was identified. Other model types were then considered to the extent they could beat the best linear regression models. Among all the models attempted, only a tuned boosting model came in with some dramatically better results. Consequently the choice in selecting the boosting model was an easy one.

ANALYSIS

Examination of the data was not performed in a single pass, but rather in an iterative way. As models evolved and insights occurred, the data was reexamined again and again.

Mailing List Classification Model

A simple logistic regression of the variables shows possible relationships (highlighted in yellow):

Predictor	VarName	Z Value
	(Intercept)	<2e-16
Region1	reg1	<2e-16
Region2	reg2	<2e-16
Region3	reg3	0.693098
Region4	reg4	0.827194
Homeowner	home	<2e-16
# Children	chld	<2e-16

Household Income	hinc	0.10838
Gender flag	genf	0.471516
Area Wealth	wrat	<2e-16
Avg Home Value	avhv	0.58301
Median Family Income	incm	0.000449
Avg Family Income	inca	0.943464
Percent Low Incm Housing	plow	0.007375
Lifetime promotions	npro	4.19E-08
Total gifts	tgif	0.031618
Largest gift	lgif	0.642623
Recent gift	rgif	0.794428
Months since last gift	tdon	5.37E-06
Months 1st/2nd gift	tlag	<2e-16
Average Gift	agif	0.643765

Examination of the data document reveals a few important facts about these data elements:

- Household income (in red) is an ordinal variable for which there is no bracketing or scaling. Intuition about basic economics suggests that charitable giving must in some way be related to income, but in the initial regression, it came up as not significant.
- Area wealth is also an ordinal variable for which there is no bracketing or scaling.
- It was odd that Regions 3 and 4 were not significant.

Since both household income and area wealth are both ordinal variables, there was no downside to turning them into categorical dummy variables like the region dummies, and it allowed for a non-linear relationship across the range of both variables. The results were stark:

- hinc2 through hinc6 were significant, with values having slight difference
- wrat2 through wrat9 were all significant, with values ranging quite a bit

Though hinc was insignificant as a numeric, the relationship became very evident as categorical dummies. The varying relationship in different wealth categories was also shown. One of the central methods of analysis was to examine cross tabulations across variables to investigate potential interactions.

For example, when you examining the regions, it becomes apparent that there the relationship varies by region. Region 2 stands alone with people that are twice as likely to donate, and Region 1 is similar to random chance, but Regions 0, 3 and 4 are twice as likely not to donate. It becomes evident as to why the two different items reg1 and reg2 came up as significant because the other three were not significantly different than a single default rate.

			Odds
	donr		0:1
Region	0	1	
Region0	523	277	2:1
Region1	362	454	1:1
Region2	436	903	1:2
Region3	314	178	2:1
Region4	354	183	2:1

Some variables have an even more dramatic relationship to the outcome. For example, having no children means that a person is six times more likely to donate than not donate, and if the person has five children, the person is fifteen times less likely to donate.

			Odds
	donr		0:1
children	0	1	
0	194	1201	1:6
1	203	201	1:1
2	770	381	2:1
3	494	162	3:1
4	237	44	5:1
5	91	6	**15:1**

Likewise, home ownership has a tremendous impact, and there is a 9:1 lower likelihood of not donating to donating for those who do not own a home.

			Odds
	donr		0:1
home	0	1	
0	417	48	**9:1**
1	1572	1947	1:1

The cross tabulations also involved bracketing of variables. Initial review of variables were at a more granular level, but inspection then allowed intelligent bracking as is shown in the months since last gift (tdon) variable. Where tdon>24 months, then the odds are 4:1 that the person is less likely to donate.

			Odds
	donr		0:1
tdon10.12	0	1	
0	1880	1908	1:1
1	109	87	1:1

tdon13.24	0	1	
0	443	182	2:1
1	1546	1813	1:1

tdon25plus	0	1	
0	1721	1924	1:1
1	268	71	4:1

Many more cross tabulations were created to allow systematic analysis of the data, specifically in the area of interactions. The following table shows how household income and children combine to have very special effects upon the odds of donation. For example, if a person is in a lower income bracket (hinc=2), and they have three or more children, they can be over twenty times less likely to not donate than to donate. One can see the relationship shift as income increases and nearly disappear by the time one is in a middle income household (hinc=4). Of course, this makes complete sense. Poorer families with more children have less household income and simply cannot afford to donate to charities. These interaction variables could valuable for the predictive model.

hinc=2	donr		
Children	0	1	Odds
0	35	115	1:1
1	40	9	4:1
2	120	16	8:1
3	71	3	24:1
4	40	2	20:1
5	16	0	16:1

hinc=3	donr		
Children	0	1	
0	18	135	1:1
1	20	15	1:1
2	83	39	2:1
3	51	14	4:1
4	26	2	13:1
5	4	0	4:1

hinc=4	donr		
Children	0	1	
0	34	642	1:1
1	58	147	1:1
2	261	267	1:1
3	151	129	1:1
4	72	35	2:1
5	33	6	6:1

Variable Selection

Because the random forest and boosting algorithms are robust against correlated variables (and indeed random forest specifically seeks to decorrelate them), these models were supplied all the variables that had been developed.

Variables for the LDA and logistic model were developed in an iterative fashion. Categorical dummies and interaction variables were created based on the cross tabulations. In addition, classification trees were run against these models to develop some additional interaction variables. Once this robust set of variables was created, a best subset algorithm was created to allow for two different operations on the LDA model. First, I could take a base model and do a best subset on up to ten new variables; validation profits was established as the measure of success. In addition I created an option that allowed one to take a mandatory base of variables, and then have a supplemental list of other variables from which up to three variables could be dropped; the selection criteria was again the validation profit. Thus, I had a basic tool for both variable inclusion as well as exclusion. This automated tool was required because there were too many variables (i.e., >50) to attempt an outright best subset selection which would have involved running tens of millions of models.

To begin, the most significant variables were placed into the mandatory base set. Then groups of variables were suggested for inclusion and the best subset process identified the best ones. This process was repeated and the variable list grew. Once profits could not be improved by adding more variables, the process was then switched to the exclusion process to remove variables which may have become a drag on performance. When I could no longer increase profits by removing variables, the process was stopped. The final list of variables was used in the LDA and QDA models. The logistic regression model was further tuned by adding a few other variables as noted.

LDA Predictor	VarName	Notes
Region1	reg1	
Region2	reg2	
Region3	reg3	Not Sig, but included
Region4	reg4	Not Sig, but included
Homeowner	home	
# Children	chld	
Household Income	hinc	Replaced
hinc categorical dummies	hinc2	
	hinc3	
	hinc4	
	hinc5	
	hinc6	
	hinc7	
Gender flag	genf	Not Included
Area Wealth	wrat	Replaced
wrat categorical dummies	wrat1	

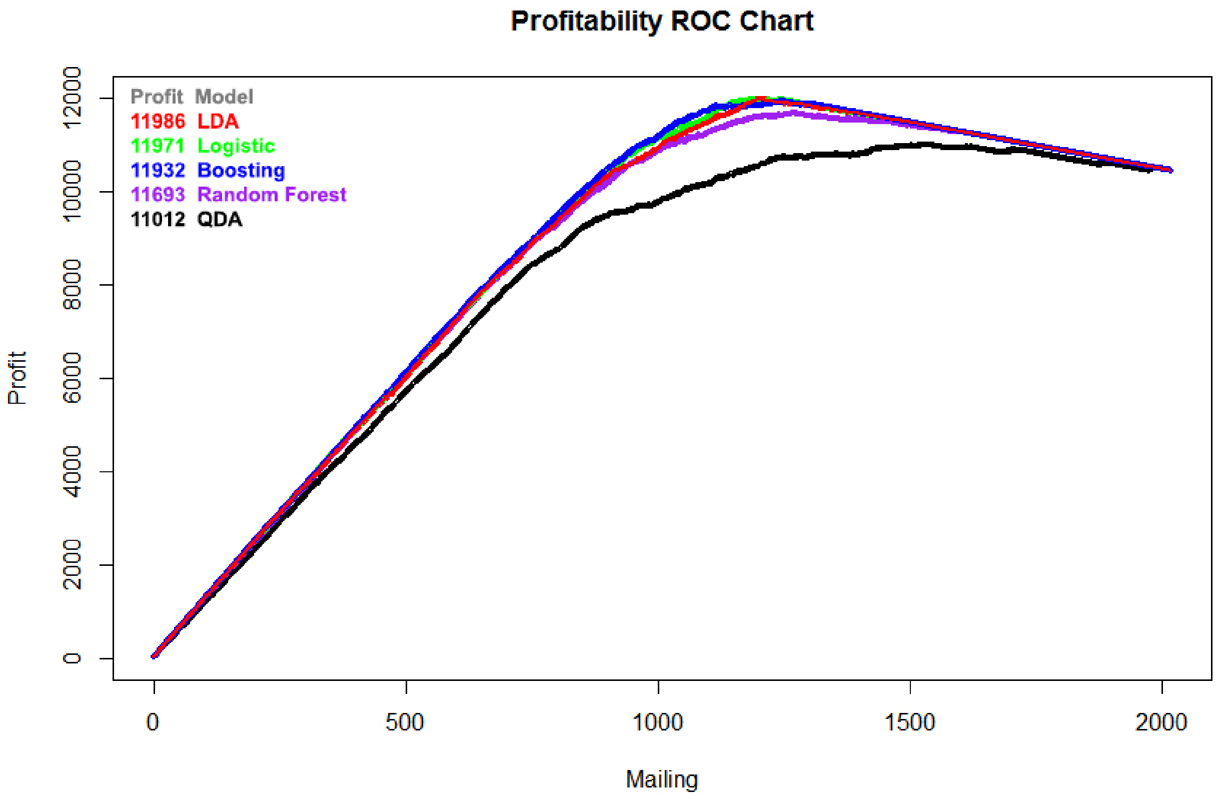
	wrat3	
	wrat4	
	wrat6	
	wrat7	
	wrat8	
	wrat9	
Avg Home Value	avhv	
Median Family Income	incm	
Avg Family Income	inca	Replaced - See ispd3
Percent Low Incm Housing	plow	Replaced-see hm1pl1
Lifetime promotions	npro	
Total gifts	Tgif	Not Included
Largest gift	Lgif	Not Included
Recent gift	Rgif	Not Included
Months since last gift	tdon	Replaced
tdon interval dummies	tdon10.12	
	tdon13.24	
	tdon25	
Months 1st/2nd gift	Tlag	
Supplemental categorical dummy for a particular segment	tlag8plus	
Average Gift	agif	Not Included
Additional Interaction Variables		
hinc=3, chld=3	i3c3	
hinc=4, chld=1	i4c1	
hinc=4, chld=2	i4c2	
hinc=4, chld=3	i4c3	
hinc=5,chld=1	i5c1	
hinc=6,chld=0	i6c0	
region=0, hinc=5	reg0hinc5	
region=2, hinc=2	reg2hinc2	
region=2, hinc=3	reg2hinc3	
region=2, hinc=4	reg2hinc4	
region=3, hinc=5	reg3hinc5	
region=4, hinc=3	reg4hinc3	
home=1, plow=0 to 9%	hm1pl1	Replaces plow
Spread between incm and inca, 30 point spread	ispd3	Replaces inca

Note: The logistic model also included wrat and genf as well as two other interaction variables: i3c2+i6c2.

As the list shows, only a few of the original variables did not contain useful information. However, to extract the useful information and exclude noise from some of them, some had to have the relevant portion of their interactions be constructed. For some, the use of additional kicker variables for specific interactions aided the models. Finally, it should be noted that the Region3 and Region4 variables were included even though they were not technically significant; I considered their inclusion to be a best practice to provide for future adjustment to the model.

Results of Classification Models

Multiple model types were developed through iteration. Only five are displayed in this chart for comparison purposes, though others were also explored: a) linear discriminant analysis; b) logistic regression; c) random forest; d) boosting; and e) QDA. The primary form of evaluation criteria within and across models was the profit performance criteria on the validation data. As you can see in the following chart for mailing profitability against the validation data, the LDA model had the highest profit:



For purposes of this project, the LDA model was been selected for submission of the results. The following observations relating to the models should be noted:

- The boosting model is particularly strong, but it required 30,000 trees and access to all of the variables that had been created. Though the training time was still in the minutes, it was a slow learner and took considerably longer than any of the other models to train. Its performance graph looks very good and natural. If the LDA were to fail in the test set, then the boosting model ought to be a second choice.

- The QDA model performed much worse than expected given the significant tuning of the LDA model parameters. Because its out-of-the-box performance was not very good, little time was invested in tuning the QDA model further.
- The random forest model did not use factors for the donr response variable, and instead used a simple regression. This allowed for easy ordering of values given the need to produce an optimum list. Even so, the performance was quite good.
- Given the tuning that had been done for the LDA model, it was not particularly surprising that the logistic model came in second place. In some respects, the green curve is a better looking performance curve than the red LDA curve, but selection was done on the basis of maximum profit. As indicated, I would select the boosting model as a second choice.
- Numerous other models were tested including support vector machine, blended models (averages of multiple models); these are not displayed for space because the performance did not come close to the LDA and logistic models.

DONATION AMOUNT MODEL

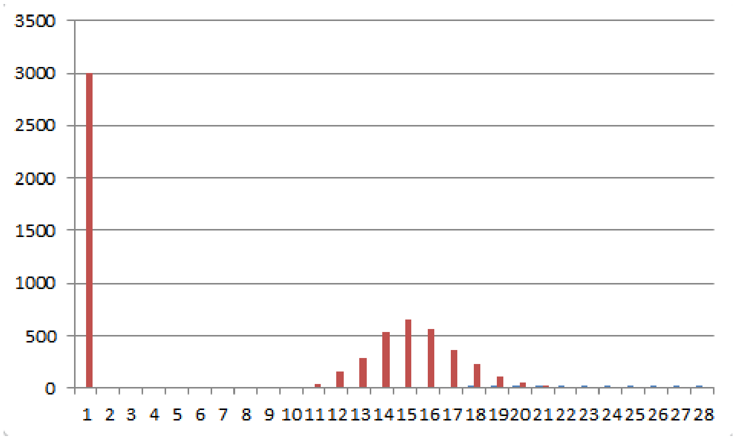
To create a reference model for the donation amount, the best subsets function was used in an iterative way to create a linear model. A group of variables was selected and the best subset was identified. Then that subset was used, along with additional variables, and the best subset identified. This iterative approach identified about 4 possible sets of variable combinations for testing. In addition, backward and forward selection identified several more for a total of seven differing sets of variables. These sets were then tested against the validation set to compute mean squared error (MSE); the results were as follows:

Model	MSE
1	1.845552
2	1.649042
3	1.660619
4	1.622105
5	1.643986
6	1.644611

The specific variable sets are not particularly important because the goal was to see what the range of MSE values were for simple linear regression and to obtain a baseline for performance. It was assumed that linear regression would not be a competitive model for final use. Of the six models generated by best subset, forward, and backward selection, the lowest MSE was 1.622.

The next model type investigated was the random forest method. Because the best value of for variable list and variable count was unknown, a best subset was run by cycling through the above list of models and a model using all available predictors (e.g., core variables, categorical dummies, interactions, etc.); different settings for the variable count were used. It came as no surprise that the best result used all predictors at a setting of 6 variables, slightly less than the square root of the predictor count. The MSE for the random forest model was 1.6265, i.e., no improvement over the linear model. The fact that the random forest model had no improvement suggested that the linear models were a reasonable baseline from which to base improvement.

The distribution of the response variable was examined with the following results:



The response variable appears to be a classic zero-inflated Poisson count distribution. Consequently, effort was taken to create a ZIP model. However, upon running a standard Poisson zip model, the MSE was 1.68, slightly above the linear model. The inability to improve performance over the linear model was a true surprise, and the code was inspected for error, but everything looked correct.

There was no point in testing a regression tree because the MSE would likely be worse than the random forest model, and so the focus moved to a boosting model, and it was with this model that some success was found. Using all available predictors, the model was first tested for interaction depth. There was no noticeable improvement in moving from 1 to 2 interaction depth, but the training time and number of trees to optimize the model increased dramatically, and so the tuning parameter of 1 was chosen. The model was then iteratively tuned for both the number of trees and the shrinkage factor; it was determined that 710 trees and a shrinkage factor of .087 gave optimal results. Here are the results of 10 runs against the validation data:

Boosting	
Run	MSE
1	1.323
2	1.311
3	1.338
4	1.320
5	1.330
6	1.332
7	1.321
8	1.322

9	1.325
10	1.328
Mean:	1.325

The MSE value of 1.31 was the lowest value observed in hundreds of iterations using many different shrinkage and tree values. This boosting model is an improvement of approximately 20% over the linear model initially derived from the best subset analysis. The results were consistent and suggest that it is nearing the optimal results achievable. Consequently, the decision to use the boosting model was an easy one to make, and the predictions in the results file use this model.

CONCLUSIONS

The ROC curves suggest that any of the top models will generate similar performance, and the top two or three models are probably within the zone where the best final test results could be driven by chance. For this submission, I selected the use of the LDA model because of its superior ROC performance. However, I have concerns that the LDA and Logistic Regression models may be overfit for the validation data. The Boosting Model would probably be a somewhat safer model to deploy because of the resiliency that boosting has to overfitting outliers. Consequently, if this were an actual mailing, I would recommend that the results of the mailing be compared to the predictions for all three top models, thereby providing guidance for any subsequent use of the models.

In this assignment, the donation amount prediction is made, but then not utilized for the mailing determination. I recommend that this prediction be used in the calculation of the cutoff. If one has the prediction, then one should place the prediction into use and decide when to stop mailing based on the likelihood of a donation combined with the likely amount of the donation. After all, there is no point to spending $12 in mailings to get an $8 donation.

In the end, it is clear that boosting is a very powerful out-of-the-box technique. I expected it to perform well for the classification problem, but was very impressed with its superior performance on the numerical prediction. This modeling method ought to be used to baseline the prediction error for any problem at the outset and then improve upon that result with another modeling technique if possible.

OPPORTUNITIES FOR IMPROVEMENT

Before and after this project was completed, I investigated opportunities for improvement over the LDA model. I speculate that there are multiple subpopulations within this single population whose signals are confusing each other in the general-purpose LDA model. To begin this investigation, the LDA model was run on the training set, and then a K-means clustering was run on the items >.15 and <.85 on the model prediction as applied to the training set. The goal was to isolate not the items whose signal was clear, but the items whose signal was conflicting with each other and to see if progress could be made on developing models to address those items.

The K-means clustering algorithm clustering started with two groups. It looked promising:

	donr=0	donr=1
Cluster 1	229	71
Cluster 2	110	698

The 2:1 grouping in favor of donors is to be expected since it is the middling group of predictions. However, things got complicated quickly, and I soon discovered that Cluster One was comprised entirely of Region 2 individuals from the same/similar income groups. Of course, my cross tabulation work had already established a number of these relationships, and it was for that reason that I had introduced interaction variables. The fact that the k-means clustering algorithm also established these relationships came as no surprise to me. Initial forays into this residual analysis several weeks ago suggests that the problem won't yield to simple pressure, especially given the fact that I have created some reasonable results by extracting some of the signal with complicated interaction variables.

I would suggest the following direction for future research for performance improvement. I would craft some automated analysis that would systematically fracture the data along different vectors (regions, income groups, wealth, children, or interaction combinations of these). These fractures would have to meet basic size criteria to provide enough data in the fractured group for clustering. Then the automation would run simple clustering (2-3 groups), and then run models on the clusters and generate aggregated profitability results. I predict that somewhere, along some fracture, there will be either a significant profitability improvement or movement towards random, and these insights might help get a better understanding of exactly what is happening within the data.

An alternative approach might be to focus all effort on the donors and ignore non-donors. This would make sense because donors only occupy 10% of the actual population, and they generate all the profit. Consequently, the goal would change from clustering on the middling populations, but instead clustering on the donr=1 population. From my cross tabulation analysis, I suspect that there are different kinds of donors: 1) the wealthy, childless donor in an urban area; 2) the poorer donor who will donate regardless of financial load; and 3) the middle-class donor who probably breaks into a couple of other categories. Those are my categories, but those are human categories based on non-comprehensive inspection.

Such analysis might prove most useful to simply identify those categories which are most profitable so that one can then begin to run experiments in marketing to identify which pitches are most productive for the most profitable category. That may be the fastest way towards higher yields than trying to eek out some minor performance from categories which are inherently bad categories.