Group Project

Critical Thinking Group 3
Joice Wong
Robert Klein
Connor Gould
Russ Conte

Northwestern University

Predict 422-DL-55
March 12, 2017

CONNECTING TO THE HEART OF CHARITABLE DONATIONS

Predicting Who will donate? How much?

TABLE OF CONTENTS

Introduction:

According to their recent mailing records, the typical overall response rate is 10%. Out of those who respond (donate) to the mailing, the average donation is $14.50. Each mailing costs $2.00 to produce and send; the mailing includes a gift of personalized address labels and assortment of cards and envelopes. It is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55.

We would like to develop a classification model using data from the most recent campaign that can effectively captures likely donors so that the expected net profit is maximized. We would also like to build a prediction model to predict expected gift amounts from donors – the data for this will consist of the records for donors only. The entire dataset consists of 3984 training observations, 2018 validation observations, and 2007 test observations. Weighted sampling has been used, over-representing the responders so that the training and validation samples have approximately equal numbers of donors and non-donors. The response rate in the test sample has the more typical 10% response rate.

Process: To do this we conducted EDA, built candidate models tested them, reported result, created a scoring file.

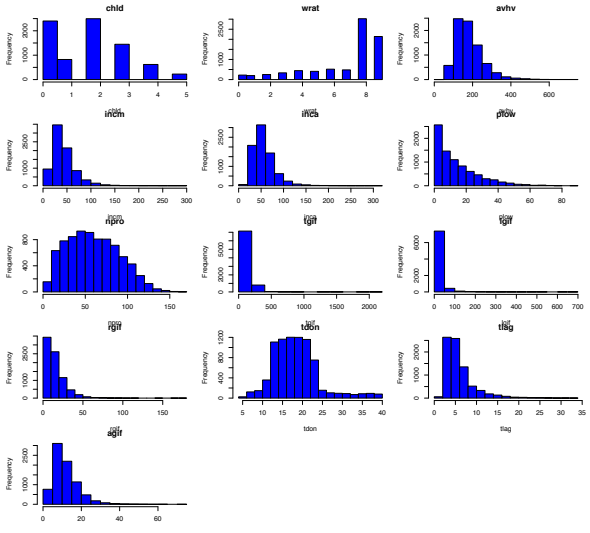
EXPLORATORY DATA ANALYSIS

Overview of EDA

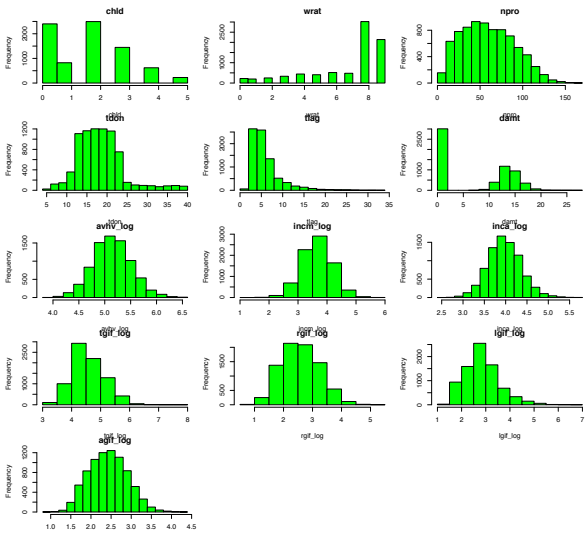
We have established two facts about our data set in our EDA. First, we will look at distributions of our data to help us understand the raw data and we have shown that all distributions are approximately normal after transformation. Second, we found that we achieved better results through transformed data, so we will show the data before and after transformation.

All distributions are approximately normal after transformation

Data before transformation



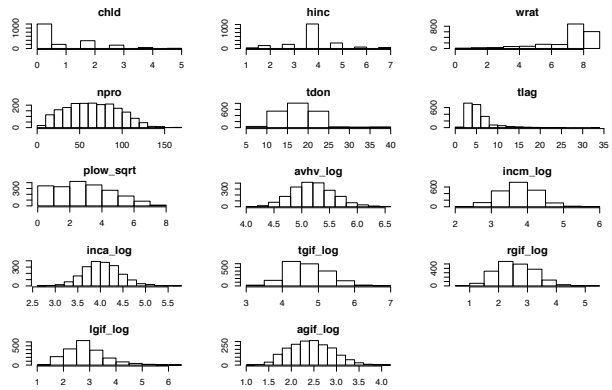
Data after transformation



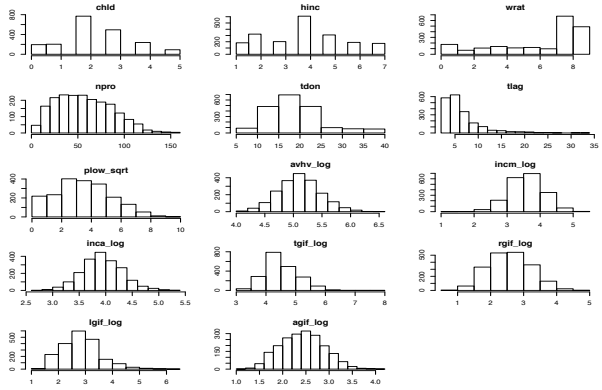
Analysis of before and after transformation: Log base e transformations were performed on only the continuous predictors. This is because we needed all of the transformations to exist in all of the data sets when we start evaluating predictors on the valid/test sets. The variables which were transformed had a distribution that is much closer to normal after the transformation. This will be critical to helping us solve the problem for the charity.

After transformation, both donor and non-donor groups are normally distributed

Donor Group

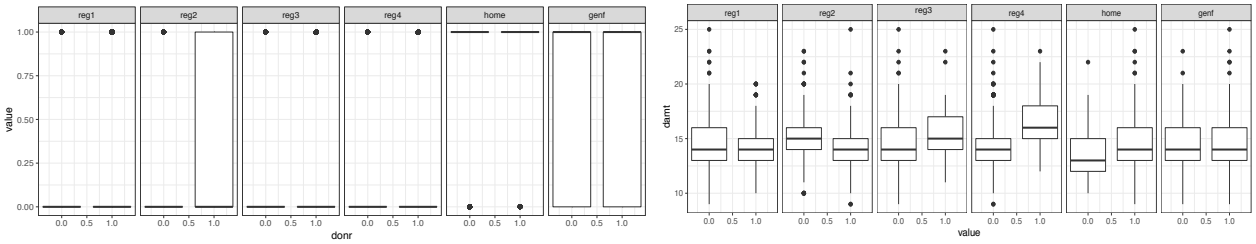


Non–Donor Group



Analysis of donors vs non–donors: As the header indicates, variables in both groups are showing normal distribution. Such fact sets a good path to linear modeling such as linear regression and linear discriminate analysis where normal distribution is an important assumption.

Box plots and bar charts of who is likely to donate

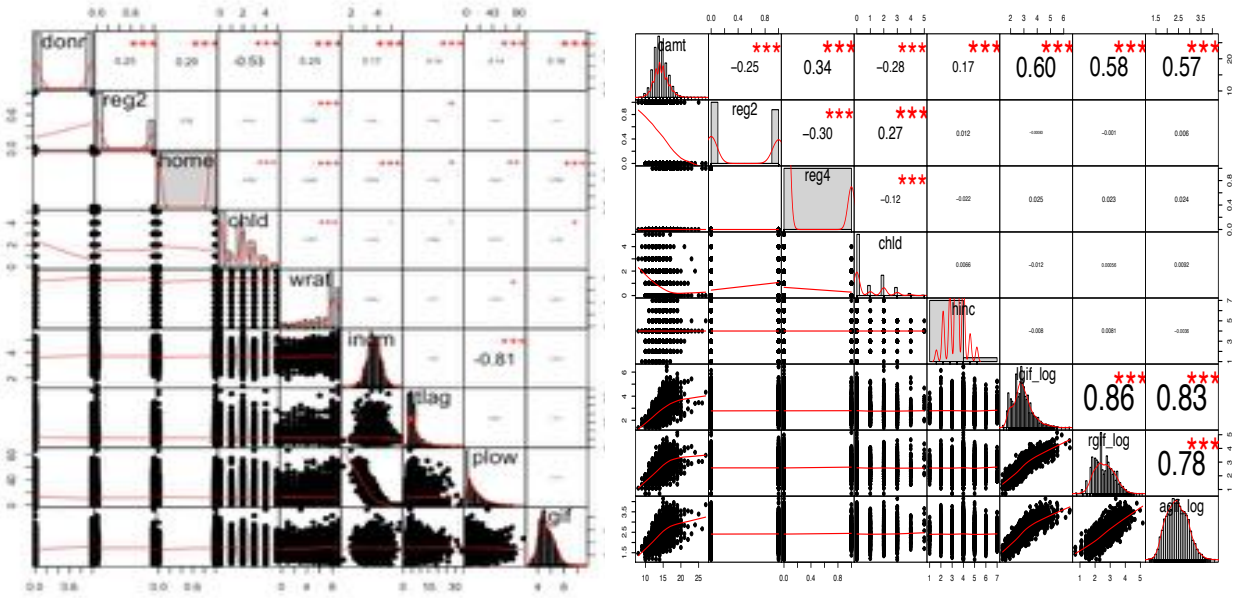


Analysis of box plots and bar charts of who is likely to donate: Since there are two response variables to be predicted in this study, a boxplot is used to identify if any of the categorical variables may be a potential good predictor and how they differ. As seen, to predict donr, reg2 is a good predictor among other variables because within group, there is an obvious difference between reg2=0 and reg2=1. On the other hand, to predict damt, reg4 is a good predictor followed by reg2, reg3 and home. It is also interesting to note that donor who lives in region 4 has a higher average donation amount.

Investigating strong predictors for Donor and Donation Amount

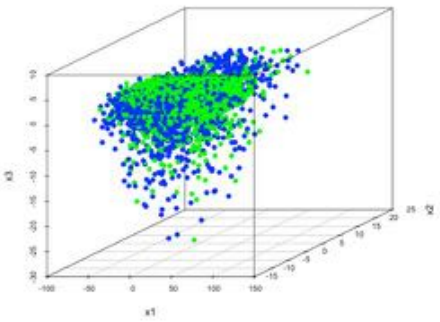
Donor

DONATION Amount



Analysis of correlation charts: The chart on the left shows the correlation with people who are likely to donate. This supports the analysis above. There is a –0.53 correlation with having at least one child, and a –0.81 correlation with income. The amount that is donated has a correlation of 0.60 with the log of gif, 0.58 with the log of rgif and 0.87 with the log of agif. In addition, the logs of agif, rgif and lgif are correlated with each other at least 0.78.

No Cluster Groups



Analysis of Significant Cluster Groups: Principal component analysis is performed and a 3D scatter plot is created. Donor group is colored in green while non-donor group is blue. Viewing from different planes, there is no obvious clustering group. It indicates that PCA/clustering analysis wouldn't work to classify donor group (at least for 3 dimensions). In addition, support vector machine cannot be used as well because no absolute line can be drawn to segregate the two groups.

The predictor variables in the training, validation, and test sets are standardized using each variables' mean and standard deviation from the training data set.

Variables	VIF
ID	1.060451
reg1	1.655486
reg2	1.958705
reg3	1.433052
reg4	1.485765
chld	1.536550
hinc	1.021033
genf	1.005185
wrat	1.082183
npro	4.472180
tdon	1.019396
tlag	1.085741
donr	40.812072
damt	41.362378
avhv_log	3.904035
incm_log	4.784119
inca_log	5.636103
tgif_log	4.555409
rgif_log	3.843139
lgif_log	5.136145
agif_log	3.400870
plow_sqrt	4.668269

We may evaluate the variance inflation factors (VIF) for each of the variables in the data set to evaluate the multicollinearity between the variables. We see that the variables with the largest VIF values are the logged inca and incm variables, as well as the logged tgif, and lgif variables. We also see that the npro variable and the square root of the plow variable have relatively large VIF values. The relatively strong multicollinearity is not surprising for these variables. The inca and inca variables represent the average and median family income of a potential donor, respectively, so we would expect that these two variables behave very similarly. We would also expect that the lgif, tgif, and rgif variables behave similarly because they represent the largest, total, and most recent gifts of each potential donor, respectively.

After removing the logged inca, lgif, and tgif variables we see that the only remaining variable with a relatively large VIF value is plow_sqrt. However, the VIF is still quite low (4.5), so we choose to keep it in the data set for analysis.

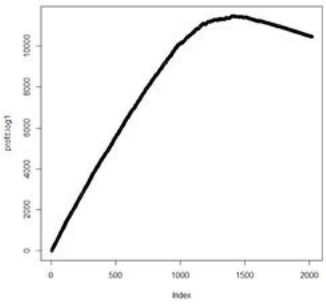
We begin the process of predicting which individuals will donate by fitting all candidate models using the training data and evaluating the fitted models using the validation data. The candidate model with the "maximum profit" is chosen as the final classification model to predict which individuals will donate. Maximum profit is calculated by first identifying the individuals in the validation data set that are likely to donate, based on the predictive model. The list of potential donors is then sorted by most-to-least likely to donate, based on the predictions from the model, and the cumulative profit is then calculated by multiplying $14.50 (the average donation) by the number of mailings to people likely to donate, and subtracting $2 (the cost of one mailing) for each mailing. We then compare the maximum profit of each candidate model and select the model with the largest maximum profit as the final classification model.

Variables	VIF
ID	1.060451
reg1	1.655486
reg2	1.958705
reg3	1.433052
reg4	1.485765
chld	1.536550
hinc	1.021033
genf	1.005185
wrat	1.082183
npro	4.472180
tdon	1.019396
tlag	1.085741
donr	40.812072
damt	41.362378
avhv_log	3.904035
incm_log	4.784119
inca_log	5.636103
tgif_log	4.555409
rgif_log	3.843139
lgif_log	5.136145
agif_log	3.400870
plow_sqrt	4.668269

Logistic Regression

The first classification model built is a logistic regression model. This model includes all the remaining variables in the training data set, as well as all two-way interaction terms between all the variables. As expected, a majority of the variables and interactions included in the model are not statistically significant at the alpha=.05 or alpha=.10 level.

Logistic Regression Model 1



The resulting model is then used on the validation data set. The model predicts which individuals in the validation data set are likely to donate based on all their attributes. We then calculate the profit for each mailing to an individual in the validation set. We may then plot the profit by the number of mailings, to see how profits change as more mailings are made. The profit curve for the logistic regression model that includes all variables and interaction terms appears as follows:
Profit is maximized when 1,401 mailings are made. This results in a maximum profit of $11,466.00.

We may also study the confusion matrix to identify the sensitivity, specificity, and accuracy of the logistic regression model against the validation data set. This model has a sensitivity of 984 / (984+15) = 98.49%, and a specificity of 602 / (602+417) = 59.07%. The accuracy of the model is (602+984) / 2018 = 78.59%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	602	15
1	417	984

Logistic Regression

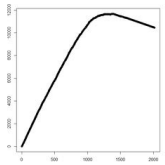
Next, we build a logistic regression model using a subset of the variables from the first logistic regression model, which included all variables and interaction terms. The subset of variables was selected by removing many of the statistically insignificant variables from the first logistic regression model. Many of the interaction terms remain in the model, particularly interaction terms including the region variables, the wealth rating, the number of children, and household income. These variables make sense intuitively to leave in the model, as one would assume that one's wealth, income, the number of children, and the area in which he or she lives would have a great impact on whether the person will donate. The variables included in the logistic regression model are as follows:

reg1	tlag	reg2:wrat	chld:npro
reg2	I(chld^2)	reg2:npro	chld:tdon
reg4	I(incm_log^2)	reg2:incm_log	chld:tlag
chld	I(hinc^2)	reg2:tlag	chld:incm_log
home	reg3	reg3:avhv_log	hinc:tdon
hinc	avhv_log	reg4:home	hinc:tlag
wrat	reg1:home	reg4:wrat	wrat:npro
incm_log	reg1:chld	chld:home	wrat:avhv_log
npro	reg1:wrat	home:npro	tlag:avhv_log
tdon	reg2:home	home:tdon	incm_log:avhv_log

The resulting model is then used on the validation data set. The model predicts which individuals in the validation data set are likely to donate based on his or her attributes that are included in the model. In this model, profit is maximized at 1,366 mailings. This results in a maximum profit of $11,666.50. The model has a sensitivity of 993 / (993+6) = 99.39%, and a specificity of 646 / (636+373) = 63.39%. The accuracy of the model is (646+993) / 2018 = 81.22%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	646	1
1	373	993

Logistic Regression Model 2

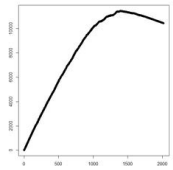


Linear Discriminant Analysis

The next model is built using linear discriminant analysis, and includes all variables and two-way interaction terms. Linear discriminant analysis often results in more stable parameter estimates than logistic regression models when the classes are separated or when the number of training observations is small and the distribution of the predictor variables are approximately normal in each of the classes. The model is fit on the training data set, and is then used on the validation data set and the results are compared to the actual values of the DONR variable in the validation data set. The maximum profit for this model is $11,461.50, which occurs at 1,396 mailings. The model has a sensitivity of 983 / (983+16) = 98.39%, and a specificity of 606 / (606+413) = 59.47%. The accuracy of the model is (606+983) / 2018 = 78.74%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	606	16
1	413	983

LDA Model #1

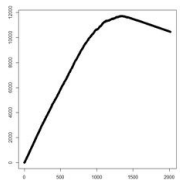


Second Linear Discriminant Analysis Model

Another linear discriminant analysis model is built, this time using the same subset of variables that was used in the second logistic regression model. This model results in a maximum profit of $11,712.50 at 1,343 mailings. The model has a sensitivity of 993 / (993+6) = 99.39%, and a specificity of 669 / (669+350) = 65.65%. The accuracy of the model is (669+993) / 2018 = 82.36%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	669	6
1	350	993

LDA Model #2

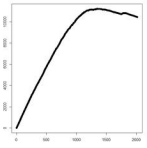


Quadratic Discrmininant Analysis

The next model is built using quadratic discriminant analysis, and includes all variables. Quadratic discriminant analysis provides a more flexible fit than linear discriminant analysis, and often outperforms discriminant analysis when the data set is very large or the assumption of a common covariance matrix for the k classes is unreasonable. The model is fit on the training data set, and is then used on the validation data set and the results are compared to the actual values of the DONR variable in the validation data set. The maximum profit for this model is $11,241.00, which occurs at 1,354 mailings. The model has a sensitivity of 962 / (962+37) = 96.29%, and a specificity of 627 / (627+392) = 61.53%. The accuracy of the model is (627+962) / 2018 = 78.74%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	627	37
1	392	962

QDA

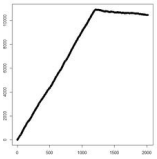


K-nearest neighbors

We now build a model using the k-nearest neighbors approach. This model identifies the five nearest observations to predict whether the individual will donate. The maximum profit for this model is $10,911.50, which occurs at 1,207 mailings. The model has a sensitivity of 919 / (919+80) = 91.99%, and a specificity of 731 / (731+288) = 71.737%. The accuracy of the model is (731+919) / 2018 = 81.76%.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	731	80
1	288	919

KNN



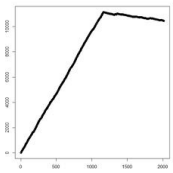
Classification Tree

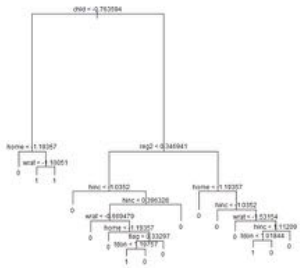
We now build a model using a classification tree. The dendogram illustrates the logic used for the splits in the classification tree. We see that the first split was done using the chld variable. Later splits are done using the home, wrat, hinc, and region variables. Not surprisingly, these variables are all included in the subset of variables used in the earlier linear discriminant analysis and logistic regression models. This further validates that these variables have very strong predictive power compared to many of the other variables included in the data set.

The maximum profit for this model is $11,149.00, which occurs at 1,168 mailings. The model has a sensitivity of 929 / (929+70) = 92.99%, and a specificity of 783 / (783+236) = 76.84%. The accuracy of the model is (783+929) / 2018 = 84.83%. As we see, the tree in easily interpreted and very intuitive. However, trees are often very non-robust, and a small change in the data can cause a large change in the final estimated tree. Therefore, we will next build a random forest, which often improves the predictive performance of the tree.

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	783	70
1	236	929

Classification Tree



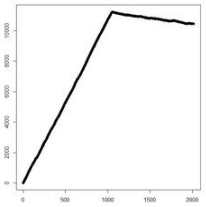


Random Forest

We now build another tree model, this time using a random forest. We allow 5 randomly selected variables to be chosen as split candidates at each new split in the tree, and any the variables in the data set can be one of the 5 options at each split. The maximum profit for this model is $11,242.50, which occurs at 1,056 mailings. The model has a sensitivity of 920 / (920+79) = 92.09%, and a specificity of 887 / (887+132) = 87.05%. The accuracy of the model is (887+920) / 2018 = 89.54%.

Random Forest

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	887	79
1	132	920

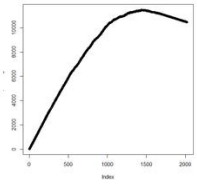


General Additive Model

The next model built is a general additive model (GAM), which provides a general framework for extending a standard linear model by allowing non-linear functions of each of the variables. The non-linear fit of GAMs can potentially make more accurate predictions, but sometimes an even more flexible approach is required. GAMs provide a useful compromise between linear and fully non-parametric models. This GAM model includes the variables reg2, home, chld, wrat, incm_log, tlag, and tgif_log, as these variables have been shown in previous models to be the most significant predictors of the DONR response variable. The maximum profit for this model is $11,445.50, which occurs at 1,433 mailings. The model has a sensitivity of 987 / (987+12) = 98.79%, and a specificity of 573 / (573+446) = 56.23%. The accuracy of the model is (573+987) / 2018 = 77.30%.

GAM

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	573	12
1	446	987

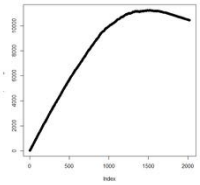


Generalized Linear Model

Finally, we build a generalized linear model. Similar to the GAM model, this GLM model includes the variables reg2, home, chld, wrat, incm_log, tlag, and tgif_log, as these variables have been shown in previous models to be the most significant predictors of the DONR response variable. The maximum profit for this model is $11,242.00, which occurs at 1,513 mailings. The model has a sensitivity of 984 / (984 + 15) = 98.49%, and a specificity of 490 / (490+529) = 48.08%. The accuracy of the model is (490+984) / 2018 = 73.04%.

GLM

Validation Set Predictions		
	Actual Value	
Predicted Value	0	1
0	490	15
1	529	984



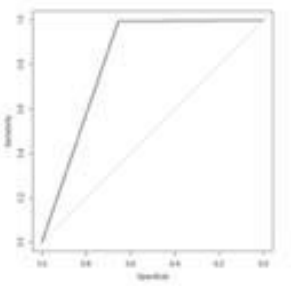
Classification Model Summary

The following table summarizes the number of mailings, maximum profit, sensitivity, specificity, and overall accuracy of each model when compared to the validation set. We see that the k-nearest neighbors model results in the smallest maximum profit ($10,911.50), and the GLM has the smallest overall accuracy (73.04%). Meanwhile, the random forest has the greatest overall accuracy (89.54%) and specificity (87.05%). However, the second linear discriminant analysis model, which includes a subset of the variables in the data, has the greatest sensitivity (99.39%, shared with the second logistic regression model) and maximum profit. As indicated by the sensitivity, the model does an excellent job of identifying the individuals that will donate. The specificity suggests that it does not do as well as several other models in predicting which individuals will not donate, so there are more false positives in this model than in many of the other models, but the ability to identify those who donate outweighs the number of false positives, which leads to the greatest validation set profit compared to all other models. In addition to the validation set results, we know that linear discriminant analysis often performs very well when the assumptions of a common covariance for the k classes is reasonable and the distributions of the predictor variables are approximately normal in each of the classes. The exploratory data analysis suggests that the distributions of the predictor variables are in fact approximately normal in each of the classes, therefore validating this assumption. This is reflected in the validation results, as we see that the linear discriminant analysis model performs the best on this data.

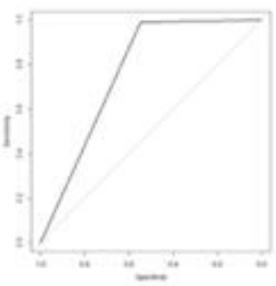
Model	Number of mailings	Profit	Sensitivity	Specificity	Accuracy	Profit rank
LDA Model #2	1,343	$11,712.50	99.39%	65.65%	82.36%	1
Logistic Regression Model #2	1,366	$11,666.50	99.39%	63.39%	81.22%	2
Logistic Regression Model #1	1,401	$11,466.00	98.49%	59.07%	78.59%	3
LDA Model #1	1,396	$11,461.50	98.39%	59.47%	78.74%	4
General Additive Model	1,433	$11,445.50	98.79%	56.23%	77.3%	5
Random Forests	1,056	$11,242.50	92.09%	87.05%	89.54%	6
GLM	1,513	$11,242.00	98.49%	48.08%	73.04%	7
QDA method	1,354	$11,241.00	96.29%	61.53%	78.74%	8
Trees	1,168	$11,149.00	92.99%	76.84%	84.83%	9
KNN	1,207	$10,911.50	91.99%	71.73%	81.76%	10

The follow plot illustrates the resulting ROC curve for the linear discriminant analysis model that was selected as the “best” model (left) and the ROC curve for a lesser-performing model, the logistic regression model using a subset of the variables (right), as the ROC curve for the linear discriminant analysis model is closer to the upper-left corner of the plot than that of the logistic regression model. The comparison shows a clear increase in prediction accuracy for the linear discriminant analysis model over the logistic regression model. The AUC of the linear discriminant analysis model (.82) also indicates the clear improvement in prediction accuracy compared to that of the logistic regression model (.76).

LDA Model #2



Logistic Regression Model #2



A PREDICTION MODEL FOR DAMT

Overview of DAMT Analysis

A number of modeling techniques and approaches were used to develop the models for the DAMT analysis including Linear Regression, Best Subset Regression, Ridge, Lasso, Partial Least Squares, Principle Component Regression, Random Forest, Boosting and Neural Nets using cross-validation approaches where applicable. The results of the analysis are divided into three parts. Data Preparation for the DAMT models, Model Build Results and then finally Model Selection.

Data Preparation for the DAMT Analysis

The data used to train and validate the models leveraged a variety of data cleansing techniques. To evaluate the effect that each of these techniques had on each of the model's performance a base line was established for the original data using Mean Squared Prediction Error and the Standard Deviation calculations. The base line results are presented in the table to the bottom.

Model Results Using the Original Data Sets

ID	Model	Mean	SD	Diff Mean	Diff SD
1	Full MLR Linear	1.88016	0.1689	0	0
2	Manually Adjusted MLR	1.86499	0.16857	0	0
3	Best Subset	1.87708	0.16969	0	0
4	Ridge	1.87128	0.17091	0	0
5	Lasso	1.85987	0.16939	0	0
6	Partial Least Squares	1.86627	0.16946	0	0
7	Principle Components Reg	2.06901	0.1834	0	0
8	Random Forest	1.6773	0.17311	0	0
9	Boosting	1.37434	0.1608	0	0
10	Neural Nets	1.92902	0.22087	0	0

Next the train and valid data sets where transformed using a log transformation on the 'avhv', 'incm', 'inca', 'tgif', 'rgif', 'lgif', 'agif' variables and a square root transformation on the 'plow' variable. This transformation generally resulted in a significant improvement as shown in the table.

Model Results Using Log transformed data sets

ID	Model	Mean	SD	Diff Mean	Diff SD
1	Full MLR Linear	1.61119	0.16082	0.26897	0.00808
2	Manually Adjusted MLR	1.6007	0.16043	0.26429	0.00814
3	Best Subset	1.61271	0.16142	0.26437	0.00827
4	Ridge	1.62053	0.16316	0.25075	0.00775
5	Lasso	1.6133	0.16176	0.24657	0.00763
6	Partial Least Squares	1.61071	0.16188	0.25556	0.00758
7	Principle Component Reg	1.71658	0.16458	0.35243	0.01882
8	Random Forest	1.66227	0.17204	0.01503	0.00107
9	Boosting	1.36763	0.1601	0.00671	0.0007
10	Neral Nets	1.54602	0.16404	0.383	0.05683

The outliers in the data were addressed through capping the 'damt', 'inca', 'lgif', 'tgif', 'avhv', 'rgif' and 'npro' variables at 1% and 99% percentiles. In addition, extreme outlier observations were dropped including observation numbers 412, 948, 343, 1579, 1505, 566, 1125, 1663, and 1069. The results again helped performance across all the models as shown in the provided table.

Model Results Using capped and transformed data sets

ID	Model	Mean	SD	Diff Mean	Diff SD
1	Full MLR Linear	1.39241	0..10355	0.21878	0.21878
2	Manually Adjusted MLR	1.37901	0.10279	0.22169	0.22169
3	Best Subset	1.38977	0.10358	0.22294	0.22294
4	Ridge	1.39383	0.10316	0.2267	0.2267
5	Lasso	1.38897	0.10346	0.22433	0.22433
6	Partial Least Squares	1.3891	0.1035	0.2209	0.2209
7	Principle Component Reg	1.48887	0.1046	0.22771	0.22771
8	Random Forest	1.3737	0.09775	0.28857	0.28857
9	Boosting	1.1277	0.09494	0.23993	0.23993
10	Neral Nets	1.24605	0.09916	0.29997	0.29997

Next the ‘Reg1’, ’Reg2’ , ’Reg3’ , ’Reg4’, ‘home’, ‘chld’, ‘hinc’, ‘wrat’ and ‘avhv_log’ variables were factorized. Generally, this greatly improved the overall models fit to the data. To the left shows the results of factorizing the above variables.

Model Results Using capped and transformed and factorized data

ID	Model	Mean	SD	Diff Mean	Diff SD
1	Full MLR Linear	1.19234	0.10035	0.20008	0.0032
2	Manually Adjusted MLR	1.18169	0.09929	0.19732	0.0035
3	Best Subset	1.18534	0.09975	0.20443	0.00383
4	Ridge	1.24985	0.10182	0.14398	0.00134
5	Lasso	1.23461	0.10184	0.15436	0.00162
6	Partial Least Squares	1.22093	0.10087	0.16888	0.00263
7	Principle Component Reg	1.46035	0.10391	0.02852	0.00069
8	Random Forest	1.37551	0.09788	-0.00181	-0.00013
9	Boosting	1.12378	0.09526	0.00392	-0.00032
10	Neral Nets	1.13781	0.09627	0.10824	0.00289

Finally a significant number of interactions between each of the variables were explored and assessed including ‘Reg1’, ’Reg2’ , ’Reg3’ , ’Reg4’, ‘home’, ‘chld’, ‘hinc’, ‘wrat’ and ‘avhv_log’. However, none of these interactions produced any significant improvements across the wide range of model types.

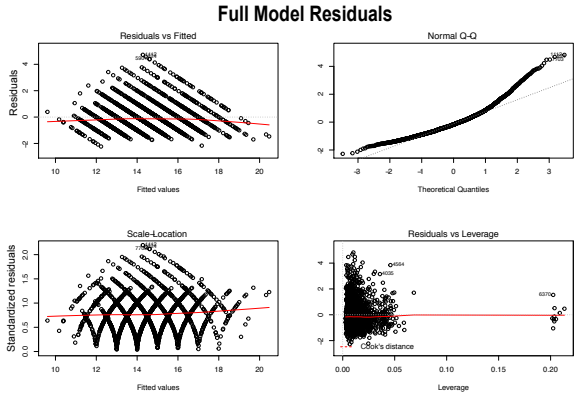
As mentioned in this section’s overview 10 different model types were built and analyzed to predict the DAMT variable. Each model used the initial training data to fit the model and then use the valid data set to measure the model’s performance by calculating the mean prediction error and standard deviation. The following provides a recap for each of these models. Please note that the PCR and Random Forest models were left out of the following analysis due to their overall non-competitive performance with the other models.

Full Linear Regression Model

lm(formula = damt.1 ~ reg1 + reg2 + reg3 + reg4 + home + chld + hinc + genf + wrat + avhv_log + sqr_incm_log + inca_log + plow_sqrt + tgif_log + lgif_log + rgif_log + agif_log, data = data.train.std.y)

A full linear regression model was built using the complete set of continuous and categorical variables contained within the data set. Using the train data set, a VIF analysis was completed and did not show any indication of multi-collinearity issues. The summary of the lm function showed most variables were statistically significant except for Reg1, Reg2, achv_log and inca_log. In addition, the Adjust R-squared was .7241 on training set. The overall p-value for the model indicated it was statistically significant at < 2.2e-16.

The model’s fit produced the residual plots shown to the right. Please note the skewness in the “QQ Plot”. The skewness is a function of the interactions between the Reg1 to Reg4 predictor variables and the DAMT response variable. The “Residuals and Leverage Plot” shows the existence of outliers mainly produced by the WRAT variable. Even though variable transformation and capping were applied to these predictor variables the residuals could not be corrected without dropping the specific variables from the model which resulted in a suboptimal performing model.

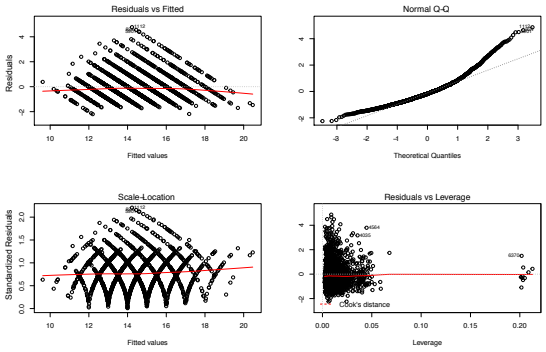


The valid data was used to generate a set of predicted damt values and then validated against a set of known damt values. This resulted in a predicted MSE for this model of 1.19234 and a standard deviation of 0.10035.

Best Subset Linear Regression Model

lm (damt.1 ~ reg3 + reg4 + home + chld + hinc + wrat + sqr_incm_log + tgif_log + rgif_log + lgif_log + agif_log + plow_sqrt, data.train.std.y)

Best Subset Linear Regression Model



A Linear Regression regsubset algorithm was ran using the training and validation datasets to identify the best predictor variable subset which resulted in the above model. Using the train data set, a VIF analysis was completed and did not show any indication of multi-collinearity issues within this model. The model fit indicated that all the subset variables were significant and produced an Adjust R-squared value of 0.723 which is slightly less than the full model. The overall p-value for the model indicated it was statistically significant at < 2.2e-16.

The model's fit produced the above residual plots. Please note the skewness in the "QQ Plot" and the "Residuals and Leverage Plot" still shows existing skewness in Reg1 to Reg4 along with the existence of outliers mainly produced by the WRAT variable.

The predicted MSE for this model using the valid data set was 1.18534 and the standard deviation of 0.09975.

Ridge Regression

For the next model a Ridge Regression model was fitted using cross-validation. A Ridge model's advantage over the linear models used above is that it attempts to improve the bias–variance trade–off by evaluating all the variables contained in the data set and reducing the uninfluential variable's coefficients to near zero in attempt to reduce noise. The model's fit results are shown below. Note that all variables have been included in the model although the coefficeints for the variables with less influence are significantly reduced to near zero. Please note that the model returned a best lambda and the lambda at SE1 of 0.1250954 and 0.2399218 respectively.

(Intercept)	5.713057
reg11	-0.20927
reg21	-0.26563
reg31	0.824161
reg41	1.619197
home1	0.954472
chld	-0.42518

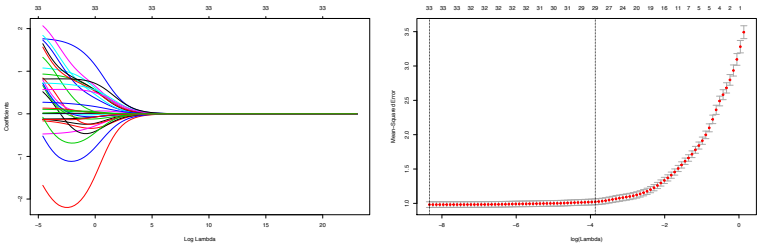
hinc2	-0.31485
hinc3	-0.02831
hinc4	0.419252
hinc5	0.812969
hinc6	0.936214
hinc7	1.164872
genf1	-0.1124

wrat1	-2.18047
wrat2	-0.68682
wrat3	-1.11601
wrat4	0.023631
wrat5	0.095193
wrat6	0.875829
wrat7	0.813886

wrat8	-0.05224
wrat9	-0.00069
npro	0.0015
tdon	0.007053
tlag	0.01244
avhv_log	-0.13702
inca_log	0.054895

tgif_log	0.23599
rgif_log	0.681544
lgif_log	0.570953
agif_log	0.816873
plow_sqrt	0.091488
sqr_incm_log	0.07794

Ridge Model Fit Results



The plots to the left are the Standardized Ridge Regression coefficients and the Mean Squared Error using the model's fit.

The MSE for this model using the valid data set was 1.2498 and the standard deviation calculated to 0.10182

Lasso Model

A Lasso model was fitted again using cross-validation option. The Lasso model is an alternative to the Ridge model above and attempts to address the Ridge model's disadvantage in including all variables into the fitted model. The Lasso model fit will actually zero out the coefficients that are identified as unimportant which makes for a simpler and more interpretable model. The results are shown below. Please note that the Lasso model produced a very similar model (32 vars) to the Ridge Model above (34 var) which indicated both model's approaches evaluated that most of the variables in the data set are statistically significant.

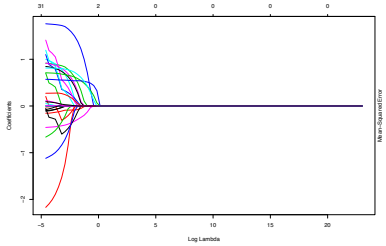
(Intercept)	4.832536
reg11	-0.10856
reg21	-0.1542
reg31	0.91805
reg41	1.759994
home1	0.999071
chld	-0.45948

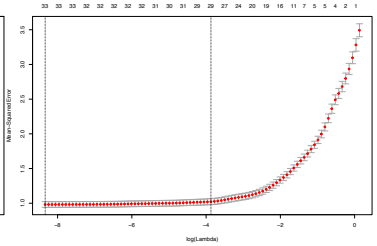
hinc2	-0.05115
hinc3	0.204218
hinc4	0.702374
hinc5	1.100575
hinc6	1.193686
hinc7	1.410125
genf1	-0.09986

wrat1	-2.17044
wrat2	-0.66575
wrat3	-1.12046
wrat4	removed
wrat5	0.052085
wrat6	0.915639
wrat7	0.846312

wrat8	-0.02173
wrat9	0.000213
npro	0.000849
tdon	0.004838
tlag	0.008343
avhv_log	-0.07793
inca_log	removed

tgif_log	0.27272
rgif_log	0.71105
lgif_log	0.571436
agif_log	0.812415
plow_sqrt	0.111705
sqr_incm_log	0.091596





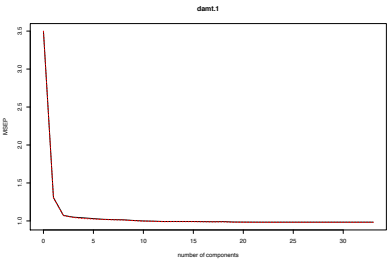
The plots to the left are the Standardized Lasso coefficients and the Mean Squared Error for the model using the model's fit. The predicted MSE using the valid data set for this model was 1.24985 and the standard deviation was calculated to 0.10182.

PLS Model

A Partial Least Squares PLS model was fitted as the other models discussed above again with the cross-validation option. PLS is a Structural Equation Modeling (SEM) technique that unlike other SEM approaches (PCR) uses supervised learning which first identifies features that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. The advantage is that the new features are related to the response variable where other SEM techniques do not. The resulting fit is shown below.

	(Intercep	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
CV	1.87	1.144	1.036	1.024	1.02	1.015	1.011	1.008	1.007
adjCV	1.87	1.143	1.035	1.023	1.016	1.013	1.01	1.007	1.006
	9 comps	**10 comps**	**11**	**12 comps**	**13 comps**	**14**	**15 comps**	**16 comps**	**17**
CV	1.003	0.9998	0.9985	0.9962	0.9965	0.9964	0.9963	0.9954	0.9947
adjCV	1.002	0.9988	0.9975	0.9953	0.9955	0.9954	0.9953	0.9944	0.9938
	18	**19 comps**	**20**	**21 comps**	**22 comps**	**23**	**24 comps**	**25 comps**	**26**
CV	0.9954	0.9929	0.9925	0.9921	0.9919	0.9919	0.9919	0.9919	0.9919
adjCV	0.9941	0.9917	0.9914	0.9911	0.9909	0.991	0.991	0.991	0.991
	27	**28 comps**	**29**	**30 comps**	**31 comps**	**32**	**33 comps**		
CV	0.9919	0.9919	0.9919	0.9919	0.9919	0.9919	0.9919		
adjCV	0.991	0.991	0.991	0.991	0.991	0.991	0.991		

PLS model Validation Plot

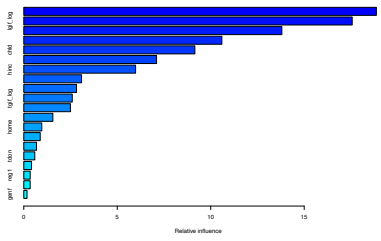


The plots to the right is the PLS validation plot which indicates there is no specific point were the cross-validation is the lowest. For prediction purposes the ncomp value was set at 12 where plot line becomes almost flat.

The predicted MSE using the valid data set for this model was 1.22093 and the standard deviation calculated to 0.10087

Boosting Model

Boosting Variable Importance



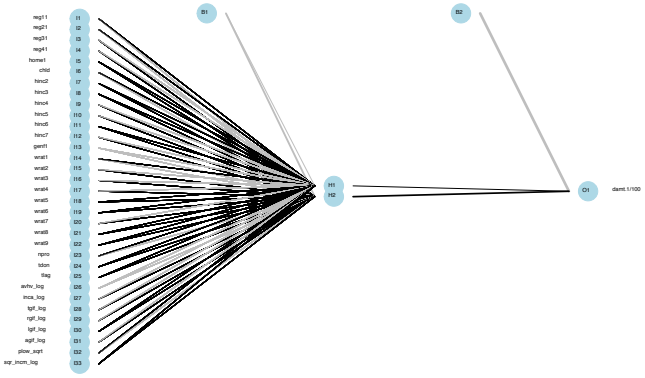
A boosting model was fitted to the transformed Charity data set as the previous models. The boosting model technique is a tree technique that involves creating multiple independent trees that are combined sequentially together to form a composite tree. This technique is advantageous to other tree methods as it learns slowly producing a tree that is less subject to variance–bias issues. The boosting model tree was grown using 1500 independent trees with the depth variable set at 3. Shrinkage for the model was set at 0.01. Below are the fit summary results which includes each variable relative's importance

var	rel.inf	var	rel.inf
rgif_log	18.86112	plow_sqrt	2.495329
lgif_log	17.56872	reg2	1.55538
agif_log	13.80552	home	0.963171
reg4	10.59665	avhv_log	0.878764
chld	9.152081	inca_log	0.681724
wrat	7.101832	tdon	0.588463
hinc	5.981533	npro	0.411143
reg3	3.089539	reg1	0.342869
sqr_incm_log	2.819566	tlag	0.341335
tgif_log	2.592632	genf	0.172634

The predicted MSE for this model using the valid data set was 1.12378 and the standard deviation calculated to 0.09526.

Neural Net Model

A Neural Net model was fitted to the transformed Charity observations leveraging both the train and valid data sets. A NNet model is a machine learning framework that attempts to mimic the learning pattern of biological neural networks. This results in a very flexible machine learning approach that can lead to improved performance over other techniques used above especially when the true relationship of the response variable to its predictors is non–linear and is highly complex in nature. Below is a pictorial of the resulting Neural Net using the NNET function fitted with the train data set.



The predicted MSE using the valid data set is 1.13781 and the standard deviation calculated to 0.09627.

As described above each of the models presented had reasonably similar results for their Mean Prediction Error except for both Random Forest and PCR. Even after standardizing and center the variables these models result did not improve. The regression models had reasonable predicted MSEs however the residual plots show the presence of skewing and outliers in the data which may be indication that the model will perform poorly. Finally, both the Neural Net and Boosting models had the lowest MSE out of the 10 DAMT models analyzed. However, the Boosting model edged out the Neural Net model which may be an indicator that the model will perform better in production.

Because the Boosting model had the best MSE, it is was selected as the champion model.

Rank	Model	MSE	SD
1	Boosting	1.12378	0.16045
2	Neural Nets	1.13781	0.09627
3	Manually Adjusted MLR	1.18169	0.09929
4	Best Subset	1.18534	0.09975
5	Full MLR Linear	1.19234	0.10035
6	Partial Least Squars	1.22093	0.10087
7	Lasso	1.23461	0.10184
8	Ridge	1.24985	0.10182
9	Random Forest	1.37551	0.09788
10	Principle Component Regression	1.46035	0.10391

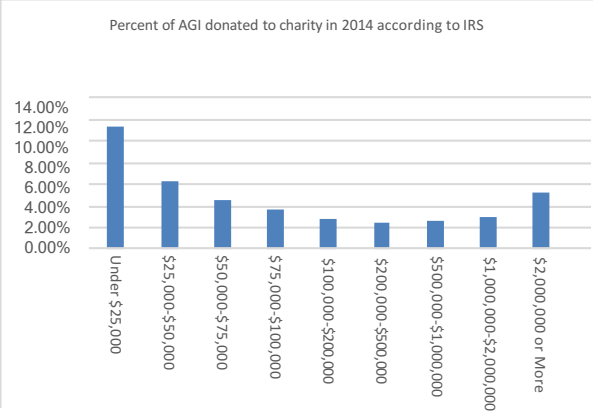
Conclusions to charity management

The chosen classification model to predict donors is a linear discriminant analysis model including a subset of the variables and interaction terms included in the data set. This model results in the largest “maximum profit” of all candidate classification models. As expected, the linear discriminant analysis performed well, which is often the case when the distributions of the predictor variables are approximately normal in each of the classes of the predictors. The chosen model to predict donation amount is a boosting model including all variables included in the data set, and using 1,500 trees, a depth of three levels, and a shrinkage parameter of .01. This model results in the smallest validation error. Boosting performs very well unless the data is noisy, because boosting models are sensitive to overfitting. However, given that the data is not very noisy, the resulting boosting model does an excellent job of predicting the donation amount. This approach recommends 1,343 mailings sent out for a profit to the charity of $11,712.50.

Next steps to consider for the charity based on the data

We noted that we only had data on one mailing and the results. It would be very useful to conduct a study across time to see what patterns may emerge.

We noted that all of the donations in this data set were under $27.00 with a minimum of $8.00 and a mean of $14.45. However, the United States Internal Revenue Service indicates that the largest number of donations come from people making less than $25,000 per year in 2014. The IRS report states the rate of donations decreases as income goes up, to around $100,000, at which point it picks back up again. The charity may be well served by broadening its range of donors to include people of all income levels. The attached chart is a summary of IRS data from 2014.



We greatly appreciate the opportunity to be of assistance, welcome feedback on our work, and wish the charity great success in their next campaign.

Sincerely,

Critical Thinking Group 3