Course Final Project

Christina Macholan | Predict 422, Section 55

INTRODUCTION

In the following analysis, we use several classification and regression machine learning techniques to build models aimed at improving the cost-effectiveness of a charitable organization's direct mail marketing campaign.

Using available records from the organization's historical donors -- which include regional indicators (reg1, reg2, reg3, reg4), demographic variables about homeownership, gender and children (home, genf, chld), economic variables about income and wealth (hinc, wrat, avhv, incm, inca, plow), number of promotions received to-date (npro), and variables about recent and past donation amounts and frequency (tgif, lgif, rgif, agif, tdon, tlag) -- we worked to find the best performing models to achieve the following:

1. A **classification model** that predicts whether or not a recipient is likely to donate (the variable "donr").
2. A **prediction model** that predicts the expected gift amounts from those who donate (the variable "damt").

ANALYSIS & RESULTS

Using a sample of 3,984 training observations, we built a set of classification models with the goal of maximizing expected net profit from the direct mailing campaign. We used the assumptions that the average donation amount is $14.50, the average campaign response rate is 10%, and a mailing costs $2.00 per item. The models were each validated against a set of 2,018 out-of-sample validation observations before the final selected model is used to predict the outcomes for 2,007 test observations.

We built the second set of prediction models to minimize the mean square error of the predicted donation amounts for donors. Again, the models were each built using a set of training data before being validated against an out-of-sample set of observations. The model that best predicted the validation data set outcomes was then used to predict the donation amounts for the test sample data.

For all models created, each variable in the training, validation, and test sets was standardized before modeling and predicting.

DATA EXPLORATION

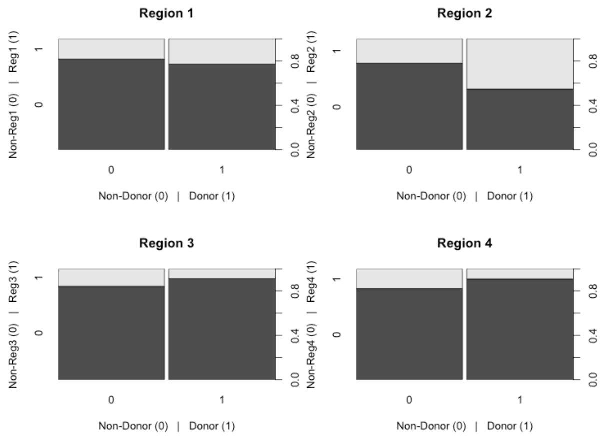
PART 1: IN-DEPTH EXPLORATORY DATA ANALYSIS

Before building our first models, we conducted a thorough analysis of the predictor variables to explore their relationship to the donor response and donation amount response variable and to confirm normality assumptions. Additionally, we looked for relationships between the predictors to identify potential collinearity issues.

During this process, several of the numerical variables were log-transformed to normalize their distribution. All variables were also standardized before the modeling process began.

Categorical Variables:

Of the four coded regions (REG1, REG2, REG3, and REG4), Region 1 shows bar for the clearest difference in donor responses, with a higher percentage of donors responding from this region than others. We should expect REG4 to be included in the classification model.



From a visualization exploration of the data, it is clear that homeownership (HOME) indicates a higher likelihood to respond to the donation mailing. We created a new flag variable for donors who have children (CHLD_YES), which shows these donors are much less likely to respond. The number of children exactly (CHLD) may also be important.

Donors with higher wealth ratings and household incomes also appear from an initial review to be more likely to donate. Gender (GENF) does not seem to have an influence on donation response.

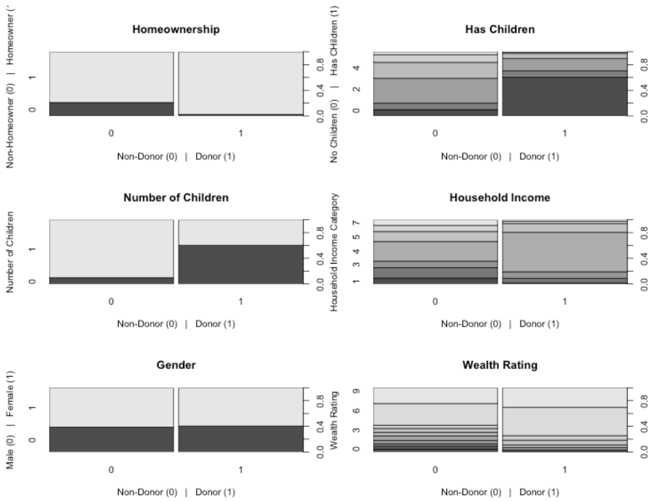
Nominal and Ordinal Variables:

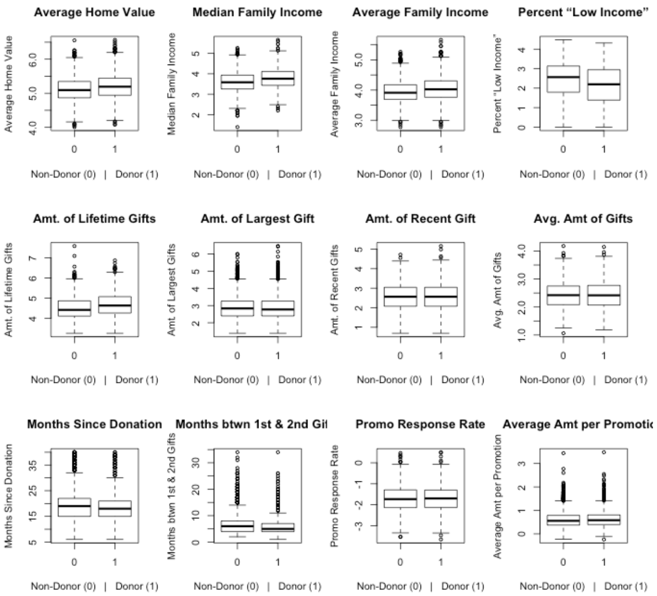
After log-transformations, each of the nominal and ordinal variables has a fairly normal distribution, as shown the boxplots to the right. Of these variables, only average home value (AVHV), the income variables (INCM, INCA) and lifetime gifts (LGIF) seem to have distinct differences between responders and non-responders.

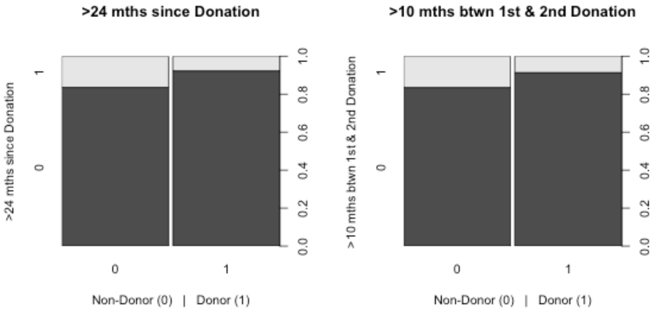
In addition to the provided variables, we chose to calculate two additional variables. One is for estimated "average response rate" (ARR), calculated from the total lifetime gifts, the average gift amount, and the number of promotions received to date. The other is for average donation amount per promotion, calculated from total lifetime gifts divided by the number of promotions received. These were included in the modeling process to determine if they had any potential predictive power.

For some of the nominal and ordinal variables "high" and "low" flags were also created as variables to include in the model (indicated by "h_" or "l_" preceding the variable name. These flags were intended to mark people with extreme donation behaviors (e.g. recent givers, largest donors, etc.)

In particular, for donors who have lapsed in the last year (24+ months since the most recent donation) and for donors who have a high time between donations (>10 months) the likelihood of donating appears to be higher.







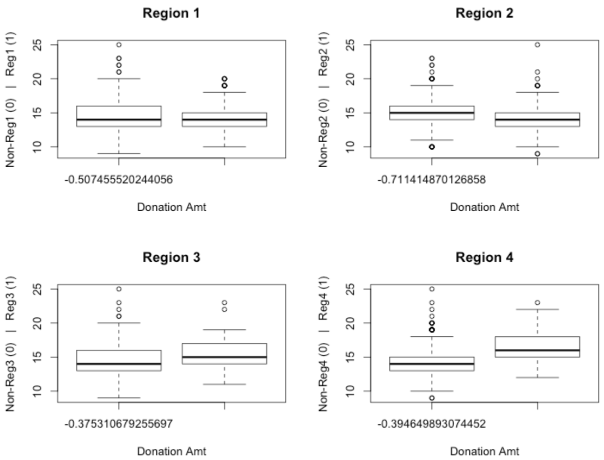
PART 2: EDA FOR DONATION AMOUNT PREDICTION

Categorical Variables:

Of the four coded region variables, Regions 3 and 4 appear to have higher average donation amounts than Regions 2 and 3. We should expect this difference to appear in our donation amount prediction model.

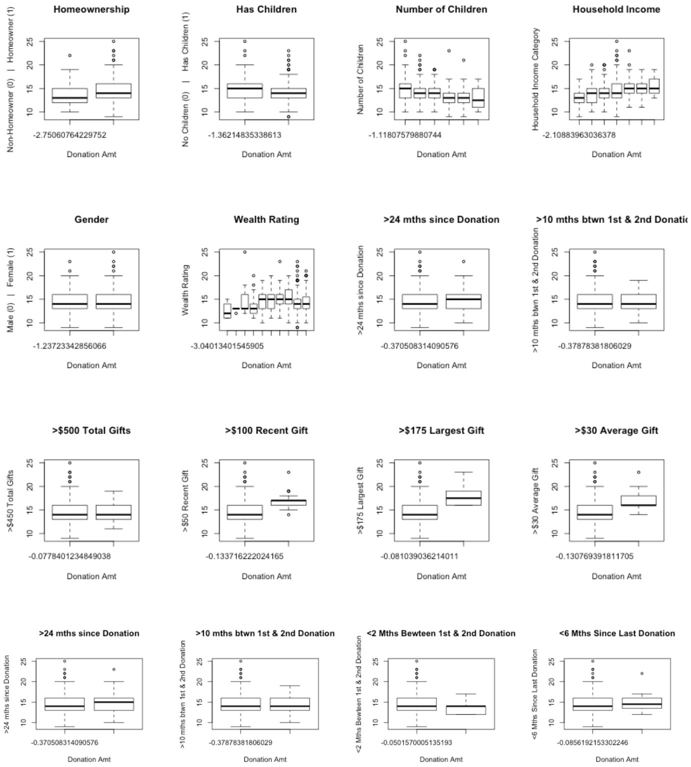
From the box plots to the right, it appears that, in general, people with more children donate less. On the other hand, homeownership, higher average and median neighborhood incomes, higher wealth ratings, and higher household income are positively associated with donation amount.

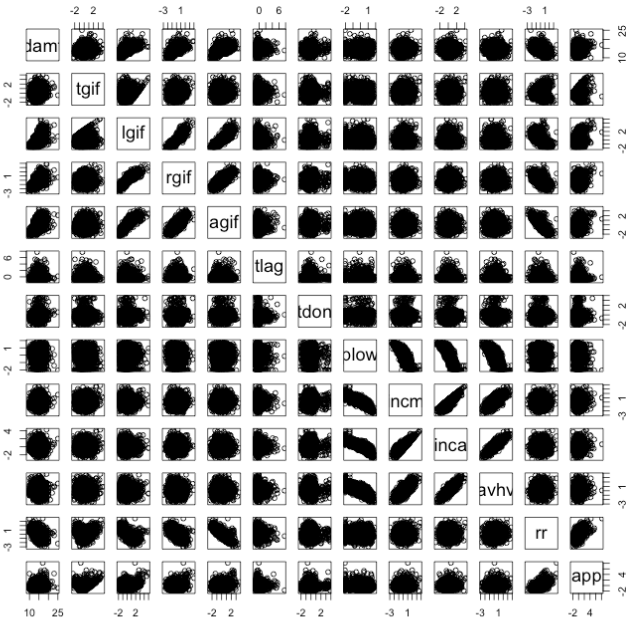
The other categorical variables initially do not show much influence on donation amount from an initial visual analysis, but some still proved to be important for prediction.



Nominal and Ordinal Variables:

A scatterplot of the nominal and ordinal variables against donation amount shows that three of the gift variables (LGIF, RGIF, AGIF) have the closest to a linear relationship with donation amount. This plot also reveals the highly collinear behavior between TGIF, LGIF, RGIF, and AGIF as well as between INCM, INCA, and AVHV. We will need to be aware of these collinearities, especially when interpreting OLS models that will be sensitive to their effects.





BUILDING CLASSIFICATION MODELS FOR DONOR RESPONSE (“DONR”)

Prior to creating any classification models, we used classification trees, the Lasso method, and best subset selection to explore the relative importance of each variable in the model. Outputs from these methods were used to select the variables used in candidate logistic regression models, LDA models, QDA models, and generalized additive models. K-nearest neighbors, random forests, and bagging and boosting methods were also explored.

LOGISTIC REGRESSION MODEL

We first used logistic regression to create a set of models including various combinations of the predictor variables. Models with relatively low expected net profit on the validation data set were discarded. The best performing logistic model was created by initially using the 19 variables selected by the Lasso method and then removing the four statistically insignificant variables (p-value > 0.05). The resulting model has an AIC of 1826.4. **Table 1** shows the model coefficient estimates for the selected logistic regression model. **Table 2** shows the results of using the model to predict the outcomes for the validation data set with 1,253 mailings and an expected maximum net profit of $11,762.00.

Table 1: Coefficients for Logistic Regression Model

Int.	reg1	reg2	home	chld	wrat	incm	npro	tgif	tlag	chld_y	h_ tdon	h_tlag	inc^2	wrat^2	chld* hinc^2
1.23	0.73	1.63	1.61	-1.12	0.62	0.79	0.30	0.39	-0.44	-1.59	-0.52	-0.27	-1.55	-0.50	-0.36

Table 2: Confusion Matrix, Number of Mailings and Expected Max. Profit from Logistic Regression Model

	TP	FP	TN	FN	N Mail	Max Profit
Model 1 *Logistic Regression 15 variables*	984	269	759	15	1,253	$11,762.00

LINEAR DISCRIMINANT ANALYSIS (LDA)

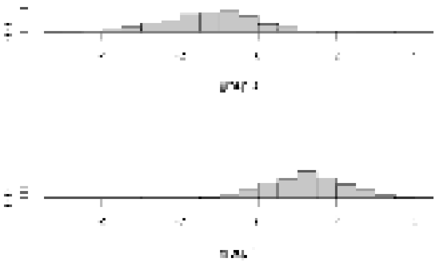
The next set of candidate models was created using the Linear Discriminant Analysis method. The method assumes normal distribution of the predictor variables with a common covariance matrix in each class. The best performing LDA model with an expected net max profit of $11,767.50 from the validation data set included all predictor variables (the 21 in the original data set as well as 17 other transformations and polynomial and interaction terms). Because the next best performing LDA model is reduced to a much simpler 19 variables and still has a high expected net maximum profit of $11,754.00, it may be preferred for prediction to avoid overfitting.

Table 3 shows the LDA coefficients for the best performing model, **Figure 1** shows the separation of the outcomes, and **Table 4** shows the results of fitting the model to the validation data set.

Table 3: Coefficients of linear discriminants:

LD1		LD1 (con’t)	
reg1	0.252421706	agif	-0.816432299
reg2	0.571167072	chld_yes	-0.611831682
reg3	-0.008283089	rr	-1.003200599
reg4	0.014821633	app	0.036457878
home	0.394131697	h_tdon	-0.183874108
chld	-0.520188316	h_tlag	-0.049129812
hinc	0.028888761	h_rgif	0.013710817
genf	-0.019493667	h_tgif	-0.021530985

Figure 1: Separation of linear discriminants:



wrat	0.243408244	h_lgif	-0.013096855
avhv	0.008234155	h_agif	0.060224848
incm	0.223817849	h_incm	-0.015068992
inca	0.014190811	h_inca	0.028191142
plow	-0.027638667	h_plow	0.031893761
npro	-0.894092953	l_tlag	-0.029616034
tgif	1.070761078	l_tdon	-0.034066615
lgif	-0.080519260	I(hinc^2)	-0.365594990
rgif	0.002334529	I(wrat^2)	-0.084222635
tdon	0.006823939	reg2:wrat	0.086091842
tlag	-0.159338175	chld:I(hinc^2)	0.068145703

Table 4: Confusion Matrix, Number of Mailings and Expected Max. Profit from LDA Model

	TP	FP	TN	FN	N Mail	Max Profit
Model 2 *LDA with all 38 predictors*	987	285	734	12	1,272	$11,767.50

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

The next set of models was created using the Quadratic Discriminant Analysis method. The method also assumes normal distribution of the predictor variables, but does not assume a common covariance matrix for the classes like LDA. Similar to the outcome we saw with the LDA approach, the best performing QDA model included all predictor variables (the 21 in the original data set as well as 17 other transformations and polynomial and interaction terms). The outcome on the validation set proposed 1,436 mailings to achieve an expected maximum profit of $11,338.00, which is much lower than any of the other models created. For this reason, QDA models were not explored further for classification.

Table 5: Confusion Matrix, Number of Mailings, and Expected Max. Profit from QDA Model

	TP	FP	TN	FN	N Mail	Max Profit
Model 3 *QDA with all 38 predictors*	980	456	563	19	1,436	$11,338.00

LOGISTIC GENERALIZED ADDITIVE MODEL (GAM)

Next, we explored a series of Generalized Additive Models. The best performing GAM proved to be the model using the 19 variables selected earlier by the Lasso method with an AIC of 1829.802. The insignificant variables ($p > 0.05$) were left in since prediction on the validation set was better with them included.

Table 6 shows the model coefficient estimates for the selected GAM with significant variables marked by an asterisk. **Table 7** shows the results of using the model to predict the outcomes for the validation data set with 1,274 mailings resulting in an expected maximum net profit of $11,778.00.

Table 6: Coefficients for Logistic GAM

Int.*	reg1*	reg2*	home*	chld*	wrat*	incm*	inca	plow	npro	tgif	tlag	chld_y	h_ tdon	h_ tlag	h_ agif	inc^2	wrat^2	reg2*wrat	chld* hinc^2
1.237	0.725	1.6134	1.601	-1.129	0.614	0.702	0.089	-0.018	0.317	0.378	-0.442	-1.562	-0.527	-0.266	0.080	-1.552	-0.493	0.1035	-0.370

Table 7: Confusion Matrix, Number of Mailings, and Expected Max. Profit from Logistic GAM

	TP	FP	TN	FN	N Mail	Max Profit
Model 4 *GAM with Lasso selected vars.*	988	286	733	11	1,274	$11,778.00

K-NEAREST NEIGHBORS

Next, we used the K-nearest neighbors algorithm to predict outcomes on the validation data set based on the "nearest neighbors" in the training data set. In order to choose the levels of *k* to optimize the performance of the model, we ran the model using values of *k* from 3 to 20 and calculated the expected maximum net profit for each level.

The best performing level of *k* was 11, predicting $11,602.00 in maximum net profit as a result of 1,159 mailings (full results in **Table 8**). Because this model did not perform better than the logistic, GAM, or LDA models, KNN was not explored further for classification.

Figure 2: Expected maximum net profit calculated for KNN with *k* from 3 to 20.

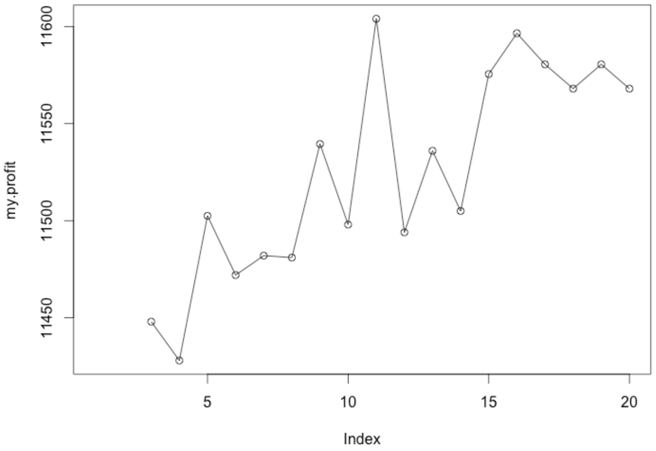


Table 8: Confusion Matrix, Number of Mailings, and Expected Max. Profit from KNN

	TP	FP	TN	FN	N Mail	Max Profit
Model 5 *k = 11*	960	199	820	39	1,159	$11,602.00

BAGGING, BOOSTING & RANDOM FORESTS

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

Between bagging and random forests, the bagged model with all variables tried at each split performed slightly better than the random forest model with only 18 variables tried at each split. Still both models performed quite poorly relative to the other classification models ($11,238.50 expected maximum net profit for the bagged model and only $11,228.00 for the random forest model). **Figure 3** and **Figure 4** show the relative importance of each variable in the bagged and random forest models, respectively.

On the other hand, the boosted decision tree models by far outperformed every other classification model that we explored. The first boosted model was able to achieve an expected maximum net profit of $11,852.50 for the validation data set. This model was created using the default 0.001 shrinkage parameter (i.e. learning rate) and 5,000 trees. As shown in **Figure 5**, the 36 variables put into the model, 26 had non-zero influence with the number of children, household income category squared and the region 2 variable carrying the most importance.

The second model was created using a separate algorithm in the "dismo" R library which uses cross-validation to select the number of trees. With a learning rate of 0.01 set, the number of trees selected via cross-validation was 2,950. With this model, we were able to achieve an expected maximum net profit of $11,910.50 on the validation data set. As shown in **Figure 6**, the 34 variables put into the model, 29 had non-zero influence with the number of children, household income category and the region 2 variable carrying the most importance.

Figure 4: Relative importance of variables for bagged model with all 36 variables (original plus calculated) at each split

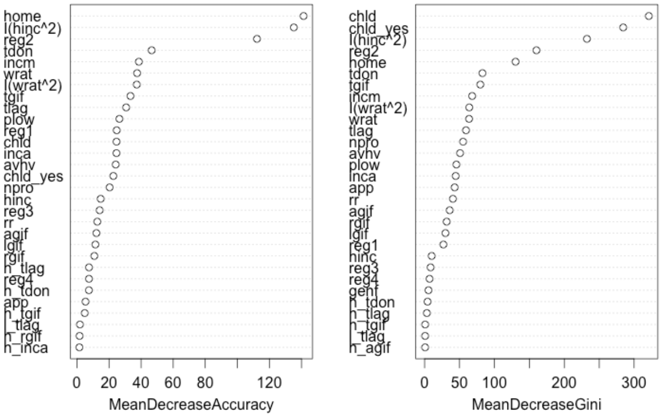


Figure 4: Relative importance of variables for Random Forest model with 18 variab
les at each split

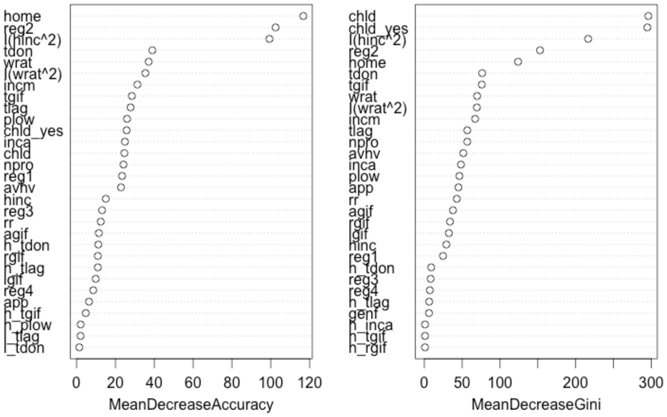


Figure 5: Relative influence plot for the Boosted model with 5,000 trees and 0.001 learning rate

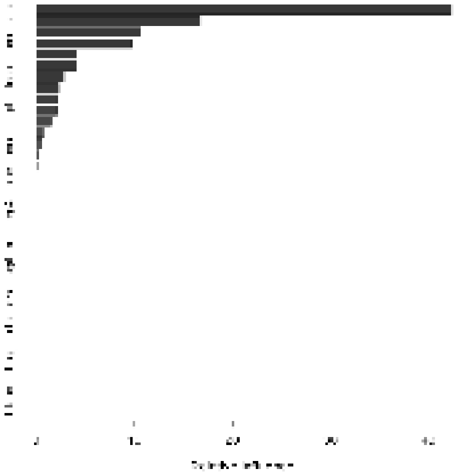


Figure 6: Relative influence plot for the Boosted model with 2,950 trees and 0.01 learning rate

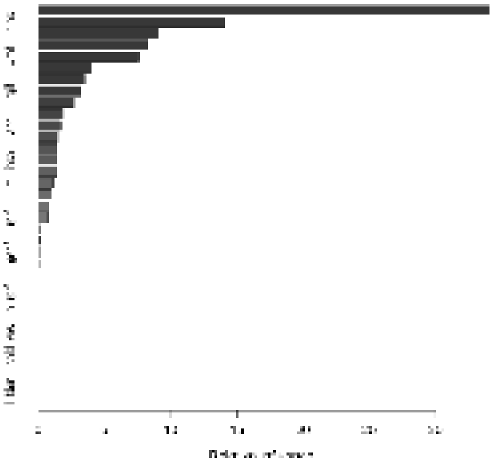


Table 8: Confusion Matrix, Number of Mailings, and Expected Max. Profit from Bagging, Boosting, & Random Forest Models

	TP	**FP**	**TN**	**FN**	**N Mail**	**Max Profit**
Model 6 *Bagging with 500 trees*	921	137	882	78	1,058	$11,238.50
Model 7 *Random Forest with 18 var split*	920	136	883	79	1,056	$11,228.00
Model 8 *Boosting with 5,000 trees & 0.001 shrinkage*	989	255	764	10	1,244	$11,852.50
Model 9 *Boosting with 2,950 trees & 0.01 shrinkage*	981	176	843	18	1,157	$11,910.50

BUILDING PREDICTION MODELS FOR DONATION AMOUNT (“DAMT”)

Prior to creating any prediction models for the donation amount, we used classification trees to explore the relative importance of each variable in the model. Outputs from these trees were used to select new interaction variables to test for inclusion (e.g. between recent gift amount or lifetime gift amount and the number of children in a household) in the Least Squares Regression, Best Subset Regression, Ridge Regression and Lasso Regression models. We also explored Principal Components Regression and Partial Least Squares Regression for variable reduction, and revisited Bagging, Boosting, and Random Forests for regression.

LEAST SQUARES REGRESSION MODEL

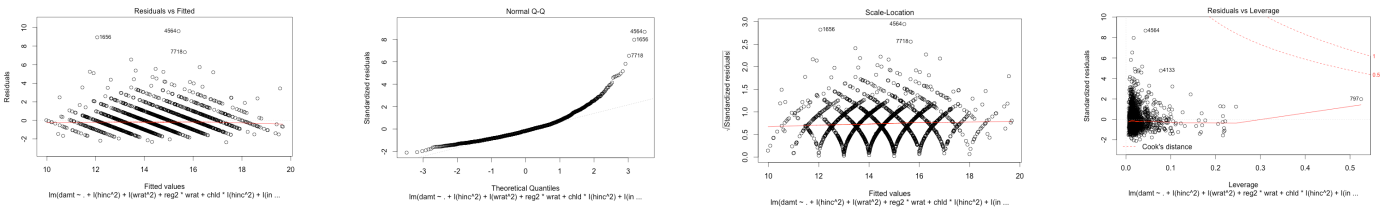
We started the donation amount prediction process by first testing the full least squares regression model, which proved to be the best performing of the Least Squares models even after limiting the candidate models to the variables selected by Lasso and the regression decision trees. This model included all of the original variables as well as the additional calculated and flag variables and some on-the-fly polynomial and interaction factor additions.

While the model has a low validation sample MSE (1.4090) and standard error (0.1530) -- which should be a reasonable estimate for the test sample MSE -- a review of the residual plots (**Figure 7**) shows non heteroscedastic distribution of the residuals and some extreme behavior of the data points in the upper quartile of the QQ plot. Because the model does not conform to the least squares assumptions, this method may be risky to employ for prediction in the long-term.

Table 9: MSE and Standard Error for Least Squares model

	N variables	Validation MSE	Std. Error
Model 1 *Least Squares Full Model*	45	1.4090	0.1530

Figure 7: Residual & QQ Plots for Least Squares Regression Model



BEST SUBSET SELECTION USING 10-FOLD CROSS-VALIDATION

Using Best Subset selection with 10-fold cross-validation, we reduced the 45 variable least squares regression model down to 16 variables without significant loss in the validation MSE (1.4153 with standard error 0.1540). This is a much simpler model that would be easier to implement and explain than the full 45 variable model. Plots showing the results of the cross-validation and variable selection are shown in **Figure 8**.

Table 10 shows the coefficient estimates for the best subset model, with the variables for Region 4, household income, number of children, lifetime gift amount, and recent gift amount weighted the highest.

Table 10: Coefficient estimates for Best Subset selection

Variable	Coefficient Estimate
(Intercept)	14.431
reg3	0.346
reg4	0.669
home	0.268
chld	-0.677
hinc	0.532

wrat	-0.300
avhv	-0.105
incm	0.210
tgif	0.215
lgif	0.546
rgif	0.427
agif	0.323
I(hinc^2)	-0.096
I(wrat^2)	-0.432
I(lgif^2)	-0.086
I(inca^2)	0.110

Figure 8: Best Subset Selection plots

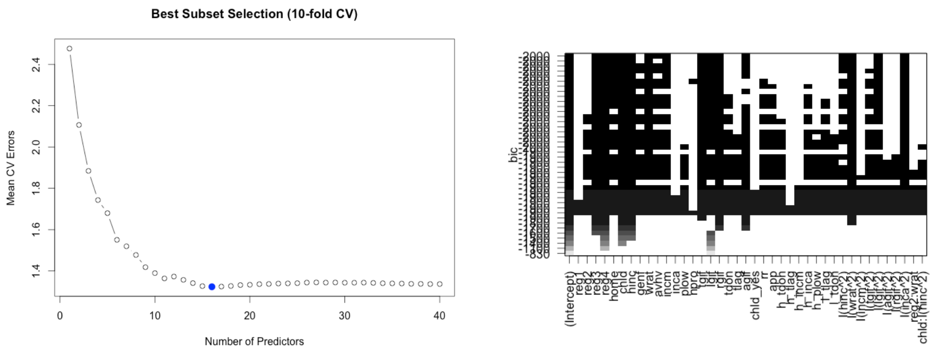


Table 11: MSE and Standard Error for the Best Subset Regression model

	N variables	Valid. MSE	Std. Error
Model 2 *Best Subset Selection with 10-fold CV*	16	1.4153	0.1540

RIDGE REGRESSION MODEL USING 10-FOLD CROSS-VALIDATION

For the next candidate models, we used ridge regression to fit a model using 10-fold cross-validation in order to select the best λ and the λ within one standard error of the minimum.

Figure 9 shows plots estimating the coefficient values and error rates for various values of λ. The coefficients in **Table 12** are for the best-λ model (λ =1.059861), which proved to have better validation MSE (1.4301) and standard error (0.1560) than the model with λ within one standard error of the minimum. Variables are highlighted to show which are the most influential (in green) and which are tuned nearly to zero (in yellow) and are expected to have little impact on the predictions.

Figure 9: Plots for Ridge Regression

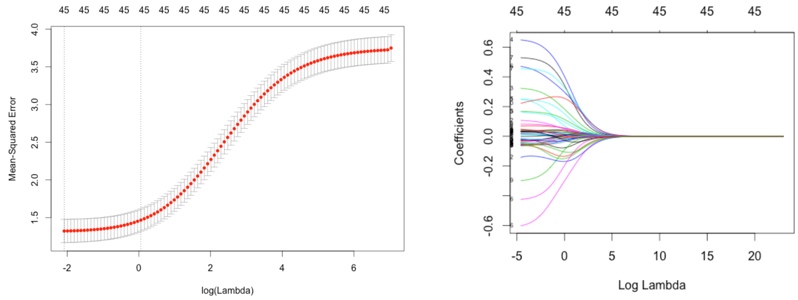


Table 12: Coefficient estimates for Ridge Regression

Variable	Coefficient Estimate
(Intercept)	14.477
reg1	-0.048
reg2	-0.093
reg3	0.285
reg4	0.593
home	0.230
chld	-0.500
hinc	0.497
genf	-0.056
wrat	-0.244
avhv	-0.052
incm	0.185
inca	-0.020
plow	0.016
npro	0.071
tgif	0.155
lgif	0.406
rgif	0.436
tdon	0.076
tlag	0.029
agif	0.259
child_yes	-0.104

Table 13: MSE and Standard Error for Ridge Regression model

	N variables	Valid. MSE	Std. Error
Model 3 *Ridge Regression with 10-fold CV* *& Best λ*	45	1.4301	0.1560

rr	-0.142
app	0.147
h_tdon	-0.038
h_tlag	0.004
h_rgif	-0.023
h_tgif	0.007
h_lgif	0.006
h_agif	0.012
h_incm	-0.021
h_inca	-0.016
h_plow	0.0407
l_tlag	0.035
l_tdon	0.028
I(hinc^2)	-0.050
I(wrat^2)	-0.371
I(incm^2)	0.039
I(tgif^2)	-0.047
I(lgif^2)	-0.038
I(agif^2)	-0.031
I(rgif^2)	-0.021
I(inca^2)	0.089
reg2:wrat	0.016
chld:I(hinc^2)	0.030
chld:lgif	-0.064

LASSO MODEL USING 10-FOLD CROSS-VALIDATION

For the next candidate models, we used lasso regression to fit a model, again using 10-fold cross-validation in order to select the best λ and the λ within one standard error of the minimum.

Figure 10 shows plots estimating the coefficient values and error rates for various values of λ. The coefficients in **Table 14** are for the best-λ model (λ =0.06973177), which proved to have better validation MSE (1.4227) and standard error (0.1550) than the model with λ within one standard error of the minimum. 8 variables were tuned out of the lasso model completely and the variables highlighted in yellow were nearly tuned nearly to zero. The variables highlighted in green are the most influential variables and overlap with the results of the ridge regression.

Figure 10: Lasso Regression Plots

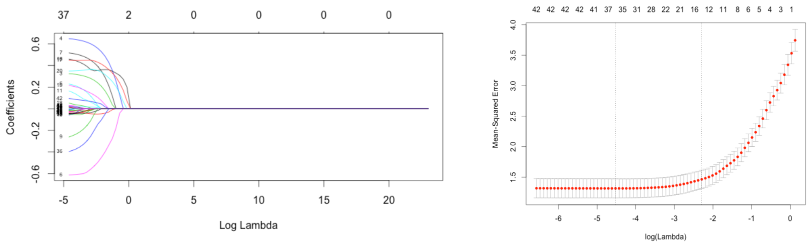


Table 14: Coefficient estimates for Lasso Regression

Variable	Coefficient Estimate
(Intercept)	14.458
reg1	-0.002
reg2	-0.038
reg3	0.323
reg4	0.646
home	0.227
chld	-0.611
hinc	0.515
genf	-0.050
wrat	-0.258
avhv	-0.057
incm	0.162
tgif	0.214
lgif	0.459
rgif	0.447

Table 15: MSE and Standard Error for Lasso Regression model

	N variables	Valid. MSE	Std. Error
Model 4 *Lasso Model with 10-fold CV*	37	1.4227	0.1550

tdon	0.051
tlag	0.018
agif	0.348
child_yes	-0.027
app	0.013
h_tdon	-0.013
h_rgif	-0.018
h_incm	-0.005
h_inca	-0.012
h_plow	0.030
l_tlag	0.024
l_tdon	0.0124
I(hinc^2)	-0.051
I(wrat^2)	-0.393
I(incm^2)	0.021
I(tgif^2)	-0.030
I(lgif^2)	-0.045
I(agif^2)	-0.022
I(rgif^2)	-0.017
I(inca^2)	0.094
reg2:wrat	0.0004
chld:I(hinc^2)	0.025
chld:lgif	-0.043

GENERALIZED ADDITIVE MODEL FOR REGRESSION

The next set of models created were a series of Generalized Additive Models. While the highest performing GAM model created included all 45 predictors with a validation set MSE of 1.4090, the model with the 19 variables selected had only a slightly higher validation set MSE (1.4358) and standard error (0.1552) and is a much simpler model to work with and explain with only a minor loss in predictability.

The variables for the 16 coefficients in the GAM model are shown in **Table 16** with reg4, chld, rgif, and wrat^2 having the highest impact on the predictions.

Table 16: Coefficient estimates for GAM

Variable	Coefficient Estimate
(Intercept)	14.448
reg2	-0.048
reg3	0.339
reg4	0.660
home	0.233
chld	-0.635
hinc	0.527
wrat	-0.304
avhv	-0.105
incm	0.114
tgif	0.198
lgif	0.385
rgif	0.473
agif	0.385
I(wrat^2)	-0.429
I(incm^2)	0.027
I(agif^2)	-0.050
I(rgif^2)	-0.055
I(inca^2)	0.087
chld:lgif	-0.055

Table 17: MSE and Standard Error for GAM

	N variables	Valid. MSE	Std. Error
Model 5 *GAM with 19 variables*	16	1.4358	0.1552

PRINCIPAL COMPONENTS REGRESSION

The next set of models was created using Principal Components regression regression, which is a dimension reduction technique that removes the covariance between the predictors. Because several of the candidate predictors in our data set have high correlations (especially the gift amount variables) this was thought to be a potentially useful solution for simplifying the prediction model.

Figure 11 shows the MSE for the training data plotted against the number of components in the PCR model. With 18 components selected, the did not perform as well as the other prediction models on the validation set (MSE = 1.6654, standard error = 0.1630).

Figure 11: MSEP plot for PCR number of components

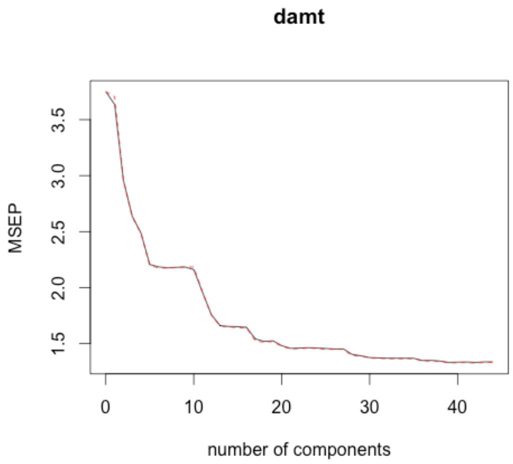


Table 18: MSE and Standard Error for PCR model

	N components	Valid. MSE	Std. Error
Model 6 *PCR with 18 components*	18	1.6654	0.1630

PARTIAL LEAST SQUARES REGRESSION

The next set of models was created using Partial Least Squares regression, which is a dimension reduction technique like PCR but that takes into account the response variable as well as the variable's relation to the other predictors.

Of the PLS models tested, the best model reduced the 45 model variables down to 7 components. **Figure 12** shows a plot of the number of components against the cross-validation mean square error for prediction with the "elbow" for this plot at 7 components and explain 63% of the variance in the training data set.

When applied to the validation data, the PLS model achieved a 1.4340 MSE with 0.1535 standard error.

Figure 12: MSEP plot for PLS number of components

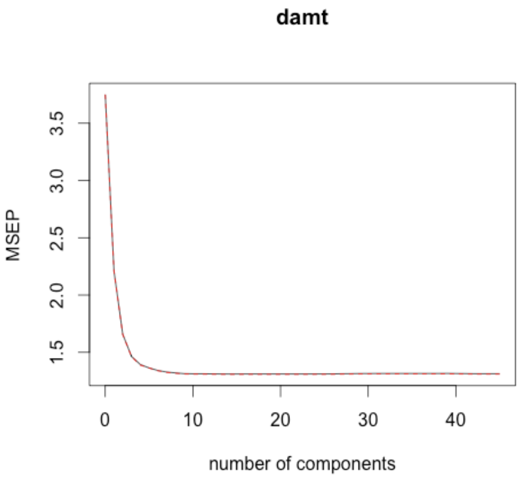


Table 19: MSE and Standard Error for PLS model

	N components	Valid. MSE	Std. Error
Model 7 *PLS with 7 components*	7	1.4340	0.1535

BAGGING, BOOSTING & RANDOM FORESTS

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

Between bagging and random forests, the bagged model with all variables tried at each split performed slightly better than the random forest model with only 18 variables tried at each split. Like with the classification problem, the bagged decision tree model performed poorly relative to other prediction models (MSE 1.7365, standard error 0.1769). **Figure 13** shows the relative importance of each variable for the bagged model.

On the other hand, the boosted decision tree model again outperformed every other classification model that we explored. The boosted model with 1,050 trees and 0.01 learning rate achieved a 1.3906 validation MSE with a standard error of 0.1769.

As shown in **Figure 14**, the 34 variables put into the model, 26 had non-zero influence with most influential variables being the recent gift amount of the donor, the lifetime gift amount, the average gift amount, whether or not they're in the fourth region, and the number of children the donor has.

This model is the final that is used for predicting the donation amounts on the test data set.

	N variables	Valid. MSE	Std. Error
Model 8 Bagging (full Random Forest)	45	1.7365	0.1769
Model 9 *Boosting with 1,050 trees & 0.01 learning rate*	26	1.3906	0.1606

Figure 13: Variable importance plots for Bagged decision tree

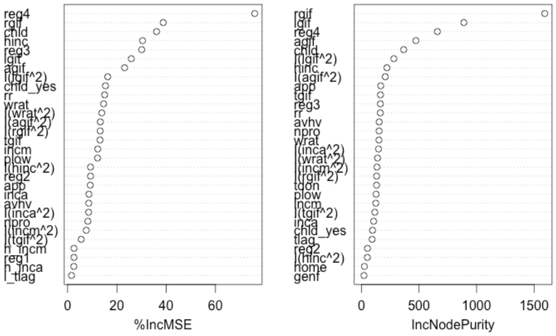
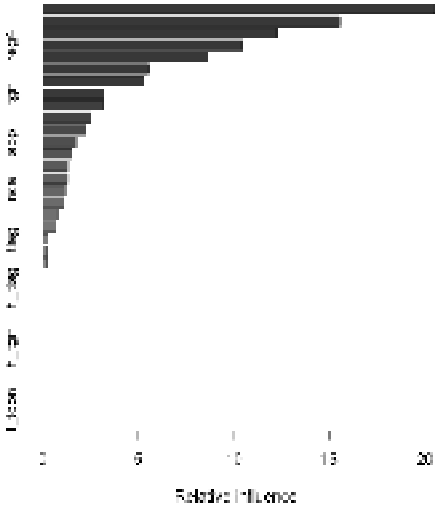


Figure 14: Variable importance plot for Boosted decision tree with 1,050 trees and 0.01 shrinkage



SELECT MODELS & PREDICT TEST DATA

CLASSIFICATION FOR DONATION RESPONSE

As was discussed above, the best performing candidate model for maximizing the expected net profit for the direct mailing campaign was the Boosted decision tree model using 2,950 trees and a 0.01 learning rate. This model achieved an $11,910.50 expected net profit on the out-of-sample validation set of data from a planned 1,157 targeted mailings. Based on the output of the classification model, we have identified 261 of the 2,007 donors in the test sample to receive the mailing.

PREDICTION FOR DONATION AMOUNT

The best model for predicted the validation data donation amounts was the Boosted model with 1,050 trees and a 0.01 learning rate. Using this model to predict the outcomes for the test data set, we have expected donation amounts that range from $9.55 to $18.63 with an average donation amount of $13.90 (slightly lower than the average amount assumed earlier).

CONCLUSION

In this analysis, we explored several statistical learning techniques to classify and predict outcomes for a charitable organization's direct mailing campaign. With a goal of improving the cost-effectiveness of the campaign, we identified one optimal model for predicting the likelihood of a mailing recipient to donate and one model for predicting the amount of donations received from donors. Of all the models explored for both classification and prediction, the boosted decision trees proved to be the best at predicting outcomes in the validation data set. A boosted model with 2,950 trees and a 0.01 learning rate predicted outcomes for the validation data set accurately enough to achieve $11,910.50 maximum expected net profit. For the donation amount prediction, a boosted model with 1,050 trees and a 0.01 learning rate achieved a 1.3906 estimated test MSE.