Table of Contents

Introduction

A charitable organization wishes to maximize its donation profits by using predictive models to determine the optimal list of donors for its mailing campaign. This course project report describes the process and techniques used to find the best performing models that will generate the optimal list of donors and give an estimate of the potential expected donation amount for these donors. It is written with sufficient detail so anyone with a reasonable statistics background can understand what has been done.

This project report describes some exploratory data analysis tasks to become familiar with these data. Data provided are split into training, validation and test – models are trained on the training data, validated on the validation data, and deployed onto test data for submission.

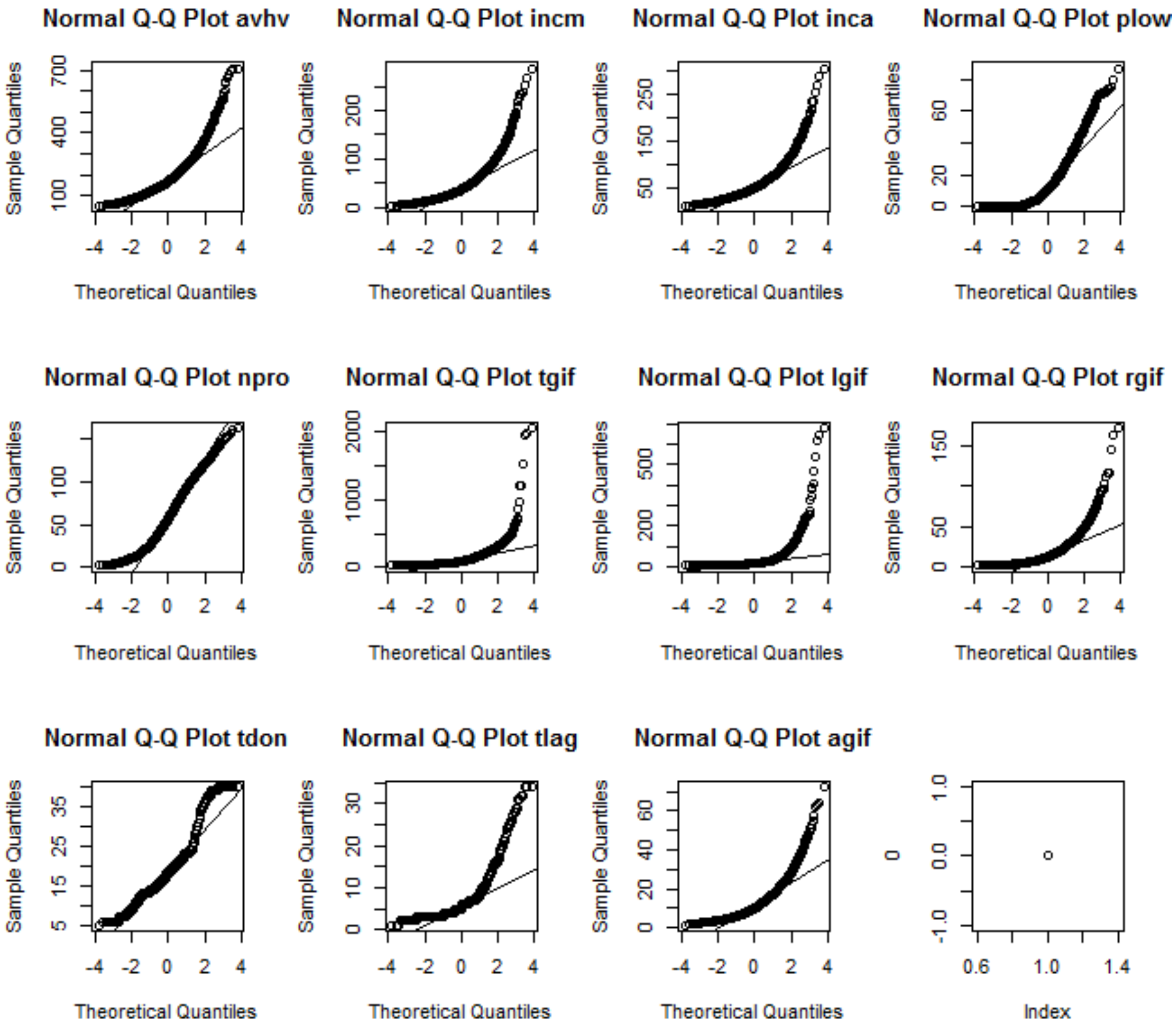
Classification models are built and validated to determine the list of likely donors. Regression models are built and validated to determine the likely donation amount for the potential donors. Different trained and validated models are shown and reasons given why the final models were chosen for deployment onto test data.

The submission of this course project includes this report, the R script that contains the detail, and the deployed results on test data containing two variables – "chat" containing 1 or 0 to indicate potential donors, and "yhat" containing the predicted donation amount if the potential donor will donate. This report provides a summary of the findings and for detail around the models the R script can be run and the output studied.

Exploratory Data Analysis

Prior to modelling, exploratory data analysis and data transforms were performed to prepare. Various techniques were employed to check for missing data and outliers. One technique, for example, was plotting normal Q-Q plots on the original data to demonstrate how each variable deviated from normal – those plots are shown in Figure 1. The predictors do not need to be normalized for most of the models tried, with the exception of least squares and discriminant analysis classification (although, discriminant analysis is more robust when predictor normality is violated and should still achieve good performance without) where the expected error may be minimized.

Figure 1: Normal Q-Q Plots on Original Data

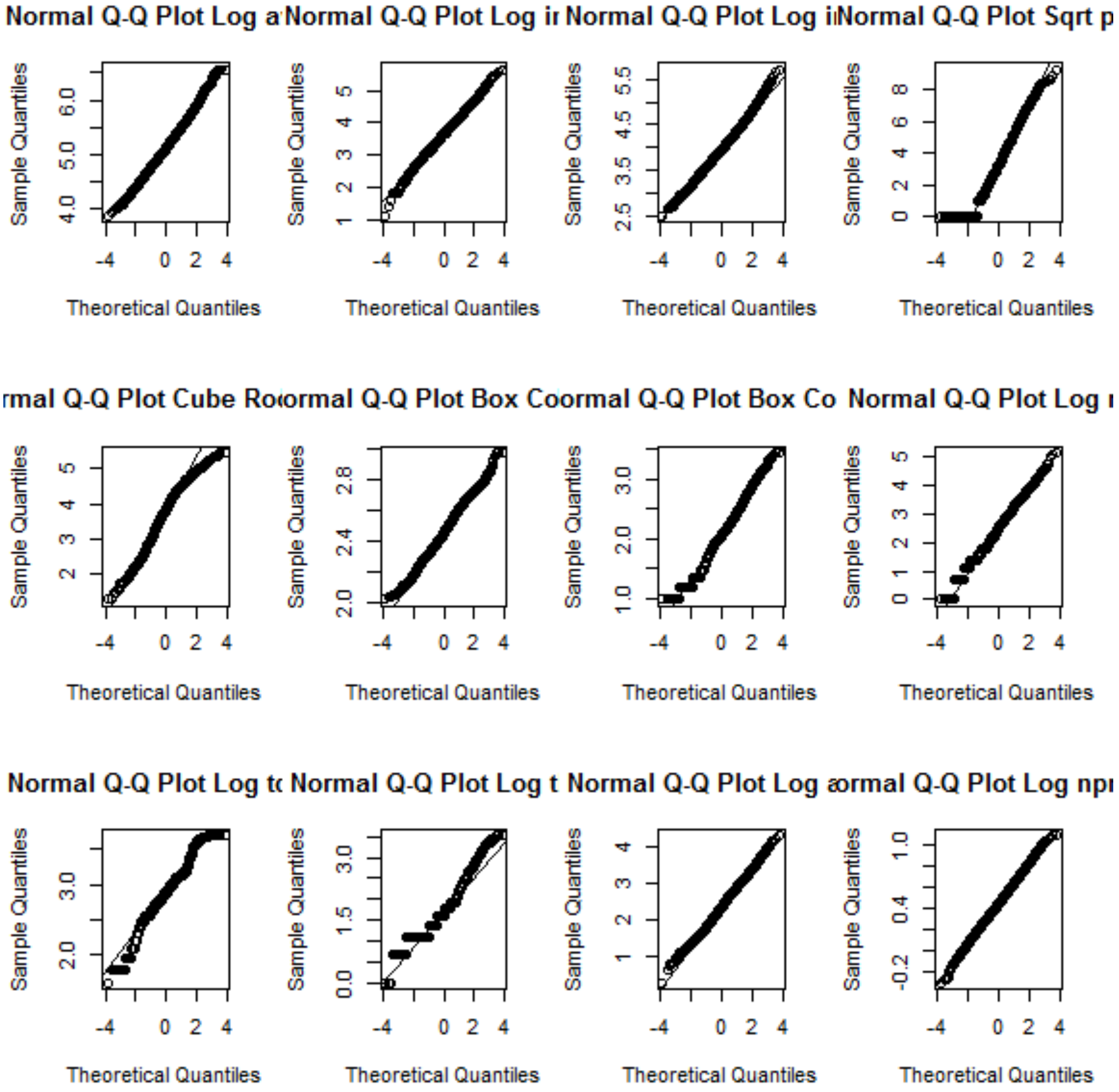


These normal Q-Q plots show the predictor variables may need to be transformed to normal distributions for some models. Most of the variables were log transformed. The exception was the plow variable which had a square root transformation. And, the tgif and lgif variables which had a Box-Cox transformation applied to each.

A correlation matrix was created between all variables and some variables were determined to be highly correlated (shown in the accompanying R script). The variables npro (number of promotions received) and tgif (dollar amount of gifts received), for example, showed high correlation (0.8). For this example, a variable was created representing the dollar amount of gifts received per promotion (tgif/npro). By using this new variable, tgif.npro, instead of using the individual variables, the issue of correlation was removed when applying to models sensitive to this (this new variable may not be used for models that don't suffer from multicollinearity issues). The resulting application of the transformations produced variables closer to normal distributions as displayed in Figure 2.

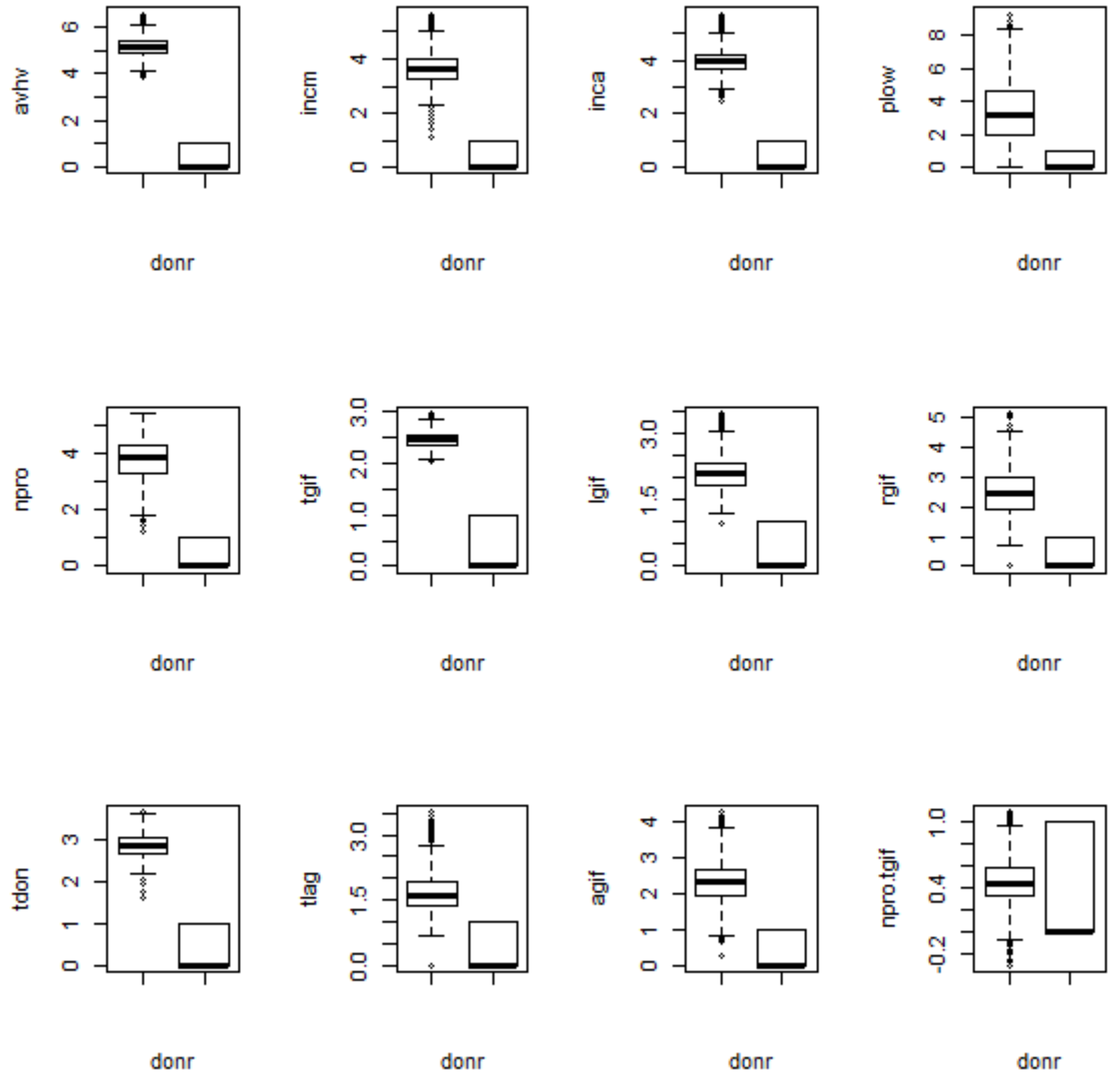
Other predictor pairs found to be highly correlated included incm/avhv (0.73), inca/incm (0.83), incm/plow (0.86), agif/lgif (0.84), and lgif/rgif (0.86). A decision was made for the relevant models where predictor correlation affects model stability (such as in linear regression models) to include one or the other variable in the pair.

Figure 2: Normal Q-Q Plots on Transformed Data



Next, the relationships between the predictor and target variables were studied. Various techniques were used, such as Box Plots, which were created for the categorical donr variable by each predictor variable, as shown in Figure 3.

Figure 3: Box Plots of Predictor Variables by DONR



These Box Plots show a clear relationship between most of these variables and the target. Linear and logistic regressions were also run with DONR and DAMT as the target variables and each individual predictor variable in order to determine whether there was a significant relationship between each predictor and its target. The predictors selected in Figure 3 show the variables with influence over the predictor DONR, with the exception of npro.tgif, which doesn't, so one of the individual variables were selected instead of the new variable for the classification models. The linear regression using npro.tgif as a predictor and DAMT as the target did show a significant correlation, so this variable was chosen for the regression models. The variables selected for initial modelling were as follows:

DONR: reg1 + reg2 + reg3 + reg4 + home + chld + hinc + wrat + avhv + inca + plow + tdon + tlag + agif + tgif

DAMT: reg1 + reg2 + reg3 + reg4 + home + chld + hinc + genf + wrat + avhv + inca + plow + rgif + tdon + tlag + agif + npro.tgif

Lasso regression was also tried for the regression models in order to generate a different list of variables. The lasso regression model takes into account the predictor variable interactions, and not just simple pairwise correlations with the individual linear regression models. A slightly different list of variables determined using lasso regression was:

DAMT: reg1 + reg2 + reg3 + reg4 + home + chld + hinc + genf + wrat + avhv + incm + plow + tgif + lgif + rgif + tdon + tlag

These predictor variables were standardized since some of the regression models will contain polynomial or interaction terms, and has the added effect of reducing multicollinearity. Other models, such as SVMs, require the predictors to be scaled so we are using the standardized predictors, otherwise the predictors containing larger magnitude would dominate the analysis.

Classification Modelling

Several binary classification models were tried using the DONR target variable (1=Donor, 0=Non-Donor). Below is a summary of each of the models tried.

Logistic Regression

A generalized linear model was built on the training data using the logit function starting with the list of variables selected during the exploratory data analysis. There were several variables that were not significant at the 0.05 level. These were removed with trial and error while attempting to also maximize the AIC. This model had the highest validation accuracy rate of 82.3% compared to the other logistic regression models attempted.

Linear Discriminant Analysis (LDA)

Two different iterations of Linear Discriminant Analysis models were attempted. The first one without interaction terms, and the second one with interaction terms. Strictly speaking, Linear Discriminant Analysis shouldn't be used on categorical targets, but can sometimes yield good predictive results.

The first model showed a validation accuracy rate of 82.4%. The second model showed a higher validation accuracy rate of 84.5%. Through trial and error and concentrating on the variables where the coefficients were far away from zero, two interaction terms were added – npro.tgif and tdon. Tdon (months since last donation) in particular had a large effect especially when adding this interaction term… the longer the time between donation, the less likely they are to donate (as seen with the negative coefficient of linear discriminants).

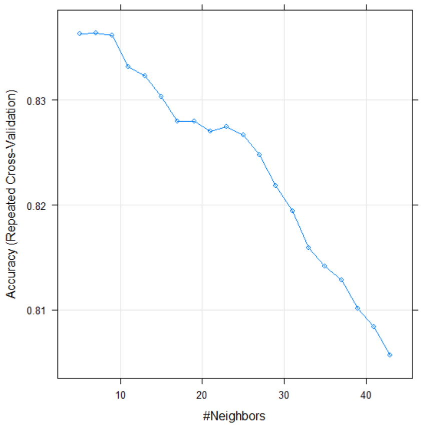
Quadratic Discriminant Analysis (QDA)

This model performed poorly with a validation accuracy rate of 73% suggesting these data are more suited to LDA with linear decision boundaries where the classes share a covariance matrix. The validation misclassification matrix showed a large number of incorrectly predicted donors (345 incorrect vs. 935 correct).

K-Nearest Neighbor (KNN) Classifier

A model that classifies a given observation to the class with the highest estimated probability, the KNN Classifier, was tried and didn't perform comparatively well with a 77.8% sensitivity accuracy rate on the validation data. Prior to building the KNN model several models were built using different values of K using repeated cross-validation. K=7 was found to be the best. A plot comparing the number of K neighbors with the accuracy rate using repeated cross-validation is shown in Figure 4.

Figure 4: Number of K Neighbors by Accuracy for KNN Model



Support Vector Machine (SVM)

Before training the SVM model, 10-fold cross validation was used to determine the optimal cost and gamma (kernel) parameters. The best performing tuning parameters produced an error of 0.158, which gave a cost of 1 (the small value gives narrow margins highly fit to the data producing higher bias and lower variance) and a gamma value for the radial kernel of 0.5.

The resulting SVM model deployed on validation data produced an accuracy of 85.7% - quite high compared to the other models attempted.

Random Forest Classification

The random forest model choses a random sample of m predictors. The m predictors selected was 5 as there were 16 predictors selected and the rule-of-thumb was used to try approximately one-third of the number of total predictors.

The resulting Random Forest model deployed on validation data produced an accuracy sensitivity of 87.0% - the highest accuracy rate of all the models so far.

Gradient Boosted Trees Classification

The Gradient Boosted Trees model is an additive model building fitting several trees on the residuals of the one before. The Gradient Boosted Trees model was found to have the highest accuracy rate compared to all the other models, so a few different models were tried and more time was spent on them to find the optimal settings.

The three models shown all use a shrinkage parameter (learning rate) of 0.01. A value of 0.1 was attempted but gave poorer accuracy on the validation set.

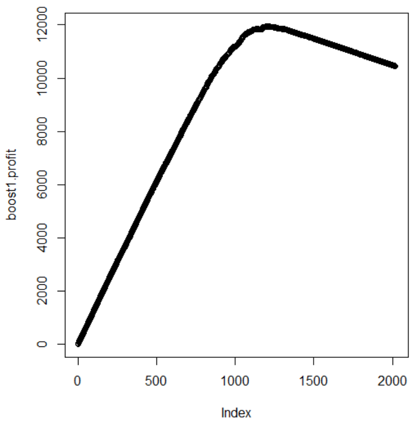
The first Gradient Boosted Trees model used a shrinkage parameter of 0.01 and interaction depth of 1. Five-fold cross-validation was used to determine the optimal number of trees; which was determined to be 7978. Other learning rates of 0.1 were tried, but gave poorer accuracy results. All variables were used in this first model, which gave an accuracy rate on validation data of 89.6%.

The second Gradient Boosted Trees model used the same settings as the first but with a much reduced set of variables. The accuracy rate on the validation set of 84.6% shows reducing the variables removed many of the predictive influencers, even though the model was simplified.

The third Gradient Boosted Trees model used a set of predictors that worked well in other models. This gave a good accuracy rate on the validation data of 89.3%.

From these results, the first Gradient Boosted Trees algorithm was studied in more detail. The maximum profit from this model was calculated at $11,959 (unadjusted for over-sampling) as shown in Figure 5.

Figure 5: Profit Chart for First Boosted Trees Model



Classification Models Compared

Maximum profit is used to select the best classification model. Since the average donation is $14.50 and it costs $2.00 to produce and send the mail, with a typical response rate of 0.1 the expected profit would be -$0.55 if everyone was mailed. Each of the models are listed in Figure 5 and it shows the maximum profit could be derived from the first Gradient Boosted Trees model with a maximum profit of $11,959 by sending 1,198 mailings if each of these people gave an average of $14.50.

Figure 5: Maximum Profit and Number of Mailings for each Model

Model	Maximum Profit	Number of Mailings
Logistic Regression	$11,410.5	1465
LDA 1	$11,383.0	1486
LDA 2	$11,458.0	1405
QDA	$11,047.0	1683
KNN	$11,280.0	1494
SVM 1	$11,234.0	1256
Random Forest	$11,766.0	1251
Boosted Trees 1	$11,959.0	1198
Boosted Trees 2	$11,574.5	1354
Boosted Trees 3	$11,918.5	1211

Applying Selected Model to Test Data

The typical response rate to mailings is 10%. However, the validation set has a higher proportion of responses – there are 999 donors vs. 1019 non-donors. This is a proportion of 0.495 = (999/(999+1019)). The results from applying the selected model to the test data are adjusted to give a scaled proportion of 0.142, which is multiplied by the number of mailings estimated on the test set. The number of mailings that may provide maximum profit on the test set would be 285. This number of mailings is used to determine a probability cutoff for the list of potential donors to mail – it is found that people with a probability of above 0.7 should be mailed. The list of donors is found in the submitted CSV file in the "chat" variable with a value of 1.

Regression Modelling

The following regression models creates predictions of expected donation amounts for each past donor. Six different types of models were attempted and are detailed below. The best model selected was determined by the lowest mean prediction error.

Lasso Regression

The lasso regression model was tried for two reasons, 1) as a dimension reduction technique due to the lasso constraint forcing coefficient estimates to zero to select variables for other models, and 2) as a means of providing predictions. The best lambda value was determined using cross-validation as approximately 0.005. The variables selected from the prediction included: reg1 + reg2 + reg3 + reg4 +

home + chld + hinc + genf + wrat + avhv + incm + plow + tgif + lgif + rgif + tdon + tlag. The model itself produced a mean validation error of 1.59.

Best Subset Selection

Best Subset Selection for variable selection was also used to determine a list of predictor variables to try in the regression models. With a maximum number of predictor variables of 19, there were 171=(19*18)/2 combinations of predictor variables tried and the model that produced the highest R-square used the following variables: reg1 + reg2 + reg3 + reg4 + home + chld + hinc + genf + wrat + incm + plow + npro + tgif + lgif + rgif + tdon + tlag + agif + npro.tgif. This list of variables was also tried in subsequent models.

Ridge Regression

It was interesting to see whether Ridge Regression outperformed Lasso Regression using the same predictors. The results appeared to be approximately the same with a mean validation error of 1.59. The best lambda value was also determined using cross-validation as approximately 0.004.

Least Squares Regression

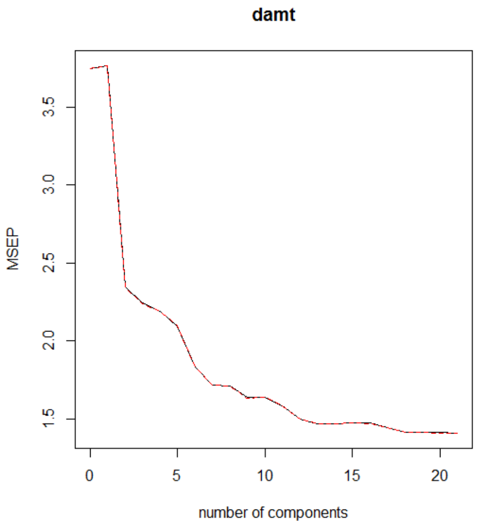
Several different Least Squares Regression models were generated using different combinations of predictor variables. Three of the models were saved in the R Script. The first model used the Best Subset Selection variables and achieved a mean prediction validation error of 1.59. However, this model created several variables that were not significant and so they were removed to give more stable results in model two, but the mean validation prediction error didn’t change at 1.59.

The third Least Squares Regression model used the variables selecting using the Lasso Regression model and then removing the insignificant predictors. This gave a worse mean validation prediction error of 1.63.

Principal Components Regression

Given the nature of this dimension reduction technique creating a set of orthogonal variables and therefore not suffering from issues of multicollinearity, and being able to increase the predictor information, all available predictor variables were used in this model. A plot showing the number of components by the root mean squared error was created to determine the number of components to use in the model from cross-validation, as shown in Figure 6. The number of components chosen was 17 out of 21, as this is the point where the graph leveled off and gave the lowest cross-validation error – other similar points were tried, but 17 gave the lowest mean validation prediction error of 1.65. This number is barely fewer than the total, which is almost similar to performing Least Squares Regression.

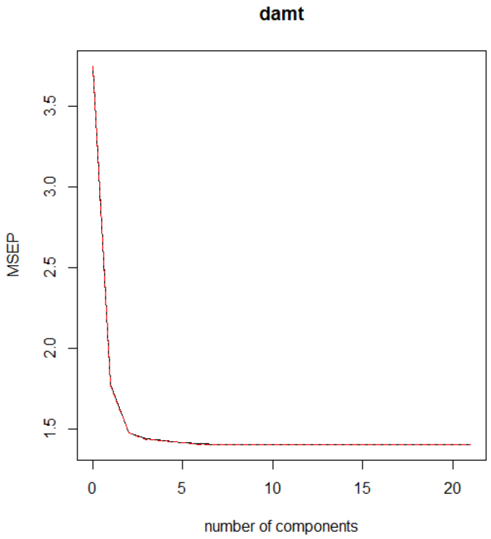
Figure 6: PCR Number of Components by Root Mean Squared Error



Partial Least Squares Regression

Partial Least Squares regression was built on all the predictors since it has similar properties to PCR. Like PCR, the models are sensitive to predictors that aren't standardized, so these are examples of where the standardization process for the predictors was useful. The mean validation prediction error obtained from a model using 7 components was 1.62. The 7 components was determined from the graph shown in Figure 7 where the line levels out, similarly obtained via cross-validation.

Figure 7: PLS Number of Components by Root Mean Squared Error



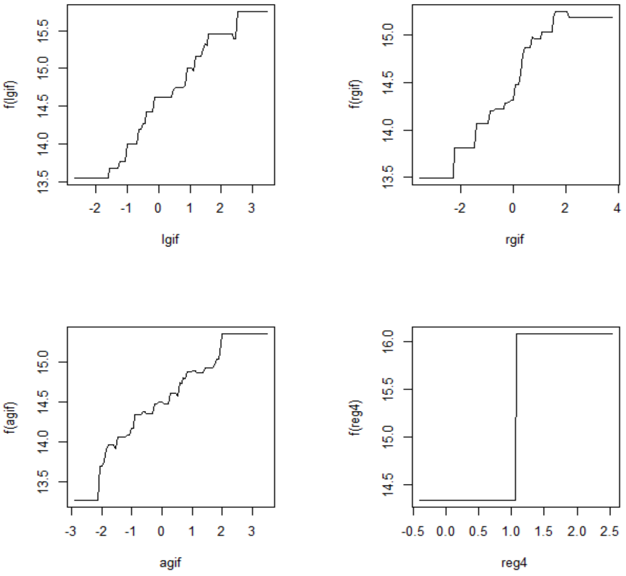
Gradient Boosted Trees Regression

Several Gradient Boosted Trees models were tried with different predictors. The models using all variables correlated to the target gave the most accurate results; these models are robust and don't suffer from multicollinearity issues found in models like OLS regression. Two models were saved in the R Script with different settings – the first model used a slower learning rate of 0.01 and the second model used a learning rate of 0.1. Both used an interaction depth of 1.

The first Gradient Boosted Trees model with a slower learning rate gave a mean validation prediction error of 1.33 and the second model gave approximately the same error. Given the fact that the faster learning rate may reduce variance, the second model should be chosen over the first. Figure 8 shows the predictor variables used in the second Gradient Boosted Trees model and the partial dependence plots for the top four predictors. These plots demonstrate the marginal effect of these predictors on the response after integrating out the other variables… as the largest dollar amount donated in the past (LGIF), the dollar amount of the most recent gift (RGIF), and the average dollar amounts to-date (AGIF) increases, so does the future predicted donation amount. Donors belonging to region 4 (REG4) also give higher donations.

Figure 8: Relative Influence of Predictors and Partial Dependence Plots for Top 4 Predictors

Predictor	**Rel.Influence**
lgif	20.6118387
rgif	15.0926423
agif	13.7665296
reg4	10.8261557
chld	8.7901978
wrat	6.8138454
hinc	5.6695603
reg3	2.9852141
tgif	2.7557816
plow	2.5189066
incm	2.3792726
reg2	1.6496335
inca	1.2722732
avhv	1.2102513
npro.tgif	0.9998995
home	0.8947136
npro	0.7600948
tdon	0.3590104
reg1	0.2573562
tlag	0.2047829
genf	0.1820400



Regression Models Compared

Similar to the classification models, the Gradient Boosted Trees Regression models outperformed the others. Figure 9 shows a breakdown of the validation set mean prediction errors and mean squared errors for each model attempted. The second Gradient Boosted Trees Regression model was ultimately selected since it has lower variance due to the faster learning rate – the mean prediction errors are virtually identical, so considered identical (1.326037 vs. 1.326595), and the second model's mean squared error is slightly lower than the first.

Figure 9: Validation Mean Prediction Error and MSE for each Regression Model

Mean Prediction Error	Mean Squared Error (MSE)	Regression Model
1.594084	0.1608267	Lasso
1.592253	0.1612298	Ridge
1.590755	0.1612799	Least Squares 1
1.592421	0.1602113	Least Squares 2
1.631829	0.1658475	Least Squares 3
1.626772	0.1594535	Principal Components
1.594573	0.1613770	Partial Least Squares
1.326037	0.1520808	Gradient Boosted Trees 1
1.326595	0.1515755	Gradient Boosted Trees 2

Conclusion

Using the charitable organizations recent mailing records, variables were prepared for analysis – checked for variable health, transformed to normality, computed from other variables, and standardized. Surprisingly, all the models performed better with normalized predictor variables – even those that are robust and don't require normalization. So, normalized and standardized predictors were used in the models.

Several classification models were built to predict potential donors and the Gradient Boosted Trees Classification model was selected giving the maximum likely profit, based on previous averages. The model identified 285 potential donors from the test set, after adjustments were made due to over-sampling.

Regression models were trained on the training data and the Gradient Boosted Trees Regression model was found to produce the lowest validation mean prediction error out of the other models. This model was used to estimate the donation amount on the test set. For the 285 potential donors recommended to mail, the average predicted donation amount is $14.36 (slightly lower than the historical $14.50 average) and is likely to raise $4,093.04 in profit.