Derek Hughes
Predict 422 – Sec 55
Spring 2015

Introduction

Marketing effectively is an extremely important goal of almost every successful company. To do so simply means one thing: spending less money to achieve more sales. Of course, improving sales without increasing expenditures would be the ultimate goal for any established company. In this case, our company is a charitable organization that has an established direct mailing campaign to receive donations that they wish to improve.

Fortunately, modern techniques in machine learning can be applied to discover insights about marketing campaigns and how to improve their effectiveness. By studying the characteristics of people who donated, we hope to discover relationships and insights that will allow us to create a machine learning model to improve the marketing campaign in terms of: 1) the ability to accurately predict whether or not a person will donate, 2) the amount of the donation, and 3) the overall profitability of the marketing campaign when the marketing expenses are included.

Analysis

General

For this study we used a data set consisting of 8009 observations from the charity's most recent marketing campaign. To do proper model creation, we split the data into three groups: a training set of 3984 observations (aprx. 50%), a validation set of 2018 observations (aprx. 25%) to test our model created on the training data, and a test set of 2007 observations (aprx. 25%) to ultimately evaluate the effectiveness of our model. This distribution ensures we have enough data to form reliable models, while still maintaining enough data on which to test our data to determine its predictive accuracy.

Variables

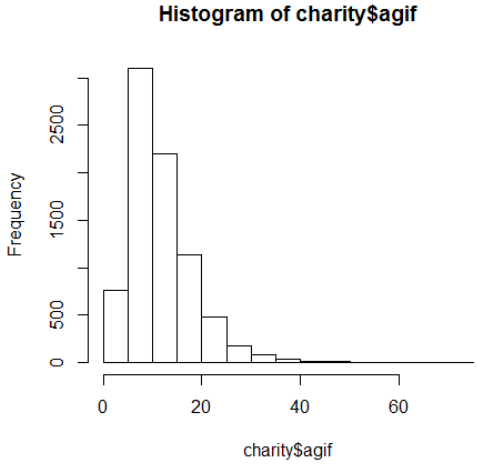
The data set consists of 24 variables. There are four variables that represent categories or segments (HOME, HINC, GENF, WRAT), one that will not be used in any model (ID), four "dummy" variables to represent five geographical regions (REG1, REG2, REG3, REG4), and two response variables (DONR, DAMT). There are two response variables because we will create a classification model to predict whether or not a person will donate (DONR) and a regression model to predict the amount of the donation (DAMT).

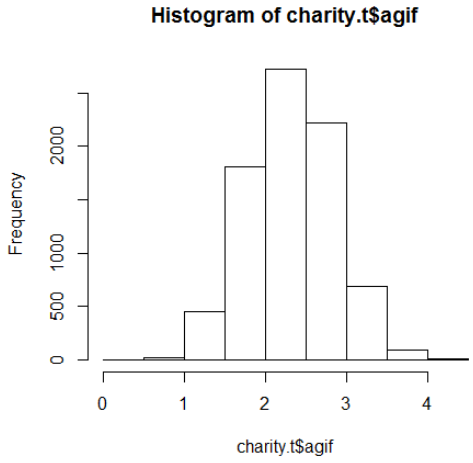
An important part of any model creation process is variable selection, imputation, and transformation, as this can improve model accuracy. We begin by running a correlation

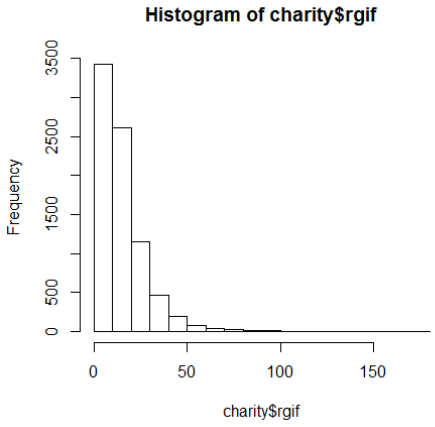
matrix over the entire group of variables to discover any strong relationships between variables. We also ran pairwise plots of the variables to visualize any correlations.

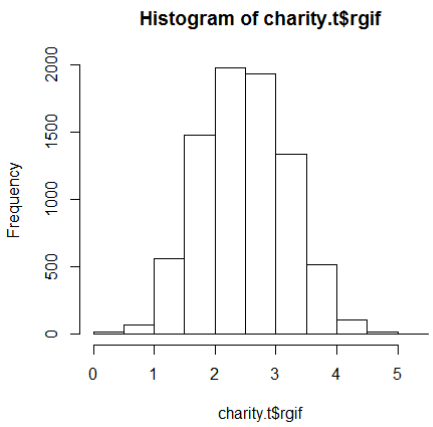
We find there are two groups of variables that have moderately strong correlations between each other (positive correlations of .62-.71 for variables AGIF-RGIF-LGIF, positive correlations of .72-.87 for variables AVHV-INCM-INCA and negative correlations of -.63 to -.65 for the variable PLOW on AVHV, INCM, and INCA). We can use this information to help us identify where we might find multicollinearity in our models; although correlations between variables does not necessary mean there will be collinearity issues in a model (VIF is a better statistic).

The next step we took was to create histograms of the frequency of each variable in the data set. We used these histograms to look for normal distributions of each variable because most models assume that each variable in the model is normally distributed. If a variable was not normally distributed, we transformed it so the transformed variable's frequency distribution more closely resembled a normal distribution. In this data set we transformed a number of variables by taking the log or square root of the variable. The variables transformed were AVHV, INCM, INCA, PLOW,TGIF, RGIF, TLAG, AGIF.









We also used the histograms and plots of the variables to identify any outliers that may exaggerate the frequency distribution curve of the variables and skew our model's results. To handle this situation we trimmed the upper values of two variables. TGIF was trimmed so any value greater than 600 was replaced with the value of 601. LGIF was trimmed to 251 for any values greater than 250. After these variables were trimmed they were transformed logarithmically as well.

Finally, we conducted some variable selection methods to attempt to simplify the model and reduce any noise created by variables that are not adding quality information into the model. We used best subset selection, forward, and backward stepwise methods to attempt to reduce the dimensions of the model. Because we are creating a classification and regression model, we ran each variable selection method with the classification variable (DONR) as the response and the regression variable (DAMT) as the response.

In every case, regardless of which variable selection method we used, the best subset, forward, and backward stepwise selected the same number and same variables. They selected 13 variables for the regression models and 9 variables for the classification models.

Variables selected for regression model:
reg1 + reg2 + reg3 + reg4 + chld + hinc + incm + plow + tgif + lgif + rgif + agif + donr

Variables selected for classification model:
reg1 + reg2 + home + chld + wrat + incm + tgif + tdon + tlag

Unfortunately as will be noted in the results section of this report, when we applied the models using either the 13 or 9 selected variables, in every model they performed worse compared to the full model or full model with some additional variables. Since the goal was to maximize the profitability of the marketing campaign, simplicity and model interpretability characteristics (which are often hallmarks of useful, practical models) were less important and may account for why the simpler models were less effective in this case.

Classification Models

For the classification models, the technique used to find the best machine learning model was to run a large number of different types of models, evaluate the results of each, and compare those results with the other models. We used eight different models: K-nearest neighbors, LDA, QDA, logistic regression, general additive models with logistic regression, random forests, boosting, and support vector machines.

The best model was determined solely by how much it maximized profitability. Profitability was determined by using the average donation amount of $14.50 and mailing

cost of $2 for each mailer. In each model, profitability increased as the number of mailings increased up to a point. After that point, profitability decreased as more mailings were sent. The maximum profitability point was the point just prior to where profitability began to decrease as more mailings were sent.

Logistic regression was used because it is the classic classification technique that has withstood the test of time. It is different from linear regression in that it models the *probability* that a response will be a given value. This prevents the problems associated with modeling binary values with a linear model and makes it ideal for classifying whether or not someone will donate to the charity. The disadvantage of the logistic regression model is that it requires each observation to be independent of the others else the model will over-emphasize those observations. It is common with research data that observations are compared or resampled over time. That means the data is not independent, and therefore can mean logistic regression is inappropriate. Furthermore, when the response variable classes are well-separated the model can be unstable. This instability is also true when n is small and the distribution of the predictors is close to normal in each of the classes.

Linear discriminant analysis (LDA) was used because it can handle those situations that often cause instability in logistic regression models. It assumes that observations are from a multivariate Gaussian distribution with a class-specific mean vector and covariance matrix common to all classes. Quadratic discriminant analysis (QDA) is very similar to LDA except that it assumes that each class uses its own covariance matrix. This improves the bias-variance trade-off over LDA.

Which model performs better depends on whether or not LDA's assumption that the classes share a common covariance matrix is met in the current data set. If not, then LDA can have high bias and QDA will generally perform better. However, LDA will generally perform better if there are fewer training observations so the reduction of variance is not as important.

K-nearest neighbors uses the average value or the most common value of a specified number of neighboring data points relative to the data point to be classified. Unlike LDA and QDA, it doesn't make any assumptions about the data and can perform quite well on more complicated decision boundaries. One disadvantage is that it doesn't provide any coefficients so we cannot tell which variables are important. However, considering that LDA and logistic regression should perform better when the decision boundary is linear, that QDA performs best with a more non-linear boundary, and that KNN performs best with an strongly non-linear boundary, by applying all four to the data set we hope to extract the best model for the data set's characteristics.

General additive models with logistic regression combine the best of linear and non-linear models. The advantage is that it allows us to apply "basis functions" that are better equipped to deal with non-linear fluctuations in the data. Hence, we can apply these functions (such as splines) to the variables that we feel are not being modeled as accurately with a linear approach. In other words, they allow us to capture information

that otherwise would not be captured. Yet, simultaneously, we can still compare the effects of the variables on the response because the model is still an additive model.

Random forests is a decision tree method which can be used to identify error rates or to classify values of a variable. It builds a number of decision tress based on bootstrapped training samples and uses a random sample of the full set of predictors (generally the square root of the number of predictors) to create a decision tree model. The point of this random use of predictors for each tree is to maximize the reduction in variability/variance of the response via de-correlation. This approach should produce a more reliable model. It should be noted that when the predictor size used in the random forest model is equal to the maximum number of predictors that the model is equivalent to bagging.

Boosting is another more recently devised approach to improving predictions based on decision tress. It can be used with classification or regression models. The big advantage of boosting over bagging and random forests is that each tree is grown using information from the previously grown trees. This allows each subsequent tree to improve on the previous by growing many small trees and combining their effects. It is a slow learning model that is controlled by a shrinkage parameter, but is often extremely powerful. Its disadvantage is that, like random forests, it can be difficult to interpret why it performs so well and is often considered a "black or magic box" model.

Support vector machines/classifier is a classification model that is often considered one of the best "out of the box" classifiers. It was originally only useful for classes separable by a linear boundary, but now can be extended to handle non-linear boundaries between classes and multiple classes. It operates by trying to select the area between classes that creates the widest boundary. How well the support vector machine can do this determines how well the model will classify the response variable values. It is distinct in the fact that it quite robust to outliers in the data, unlike other classification models.

Regression Models

For the regression models, the technique used to find the best machine learning model was to run a large number of different models, evaluate the results of each, and compare those results with the other models. We used eight different modeling techinques: best subset selection with cross-validation, principal components regression, partial least squares, least squares regression, ridge regression, lasso, random forests, and boosting.

The best model was determined solely by its mean prediction error. MPE was determined by comparing the mean difference between the predicted values from the regression model versus the response variable (DAMT) from the validation set. A lower MPE indicates the model is more accurate at predicting the response variable and, therefore, should be more useful as an accurate predictive model on new data.

The first model we deployed was a linear regression model based on the concept of least squares. Linear models are used to predict a response based on a given set of variables. By determining the difference between the predicted response by the model and the actual data, we can evaluate how well the model predicts the response we wish to evaluate. In this part of the analysis we are now predicting the amount of the donation from a donor.

One major advantage of the linear regression is that it's fairly easy to interpret the results and to identify which variables, such as largest gift donated, gender, or months since last donation, have a significant effect on predicting the amount of a donation. One major disadvantage of linear regression is that it does not produce as accurate results when the data does not have a linear relationship between the variables and the response variable.

Another model is the best subset selection model, and here we will determine the best model by using a 10-fold cross-validation method to select among models with different numbers of variables. Here we divide the training data into nine training "folds" and one test fold. We train our model on the nine folds and test it on the single fold. This process is repeated until each individual fold is used as the test fold. The MSE is then determined by averaging over all of the test folds. This process is repeated for every model size and the model size with the smallest cross-validation error is then selected. Finally, we perform a best subset selection on the full data set using the model size that was determined from the cross-validation error to perform the best. The advantage of this approach is that it provides a direct estimate of the test error and makes fewer assumptions about the true model.

The fourth model is a ridge-regression model using the same 10-fold cross-validation approach to select lambda, which is a tuning parameter that shrinks the coefficients of the variables closer to zero (or regularizes the coefficient estimates). This helps to reduce the variance of the coefficient estimates significantly which allows us to have a more stable and accurate model. Using 10-fold cross-validation to select lambda is important because a larger lambda value means the shrinkage penalty is larger. This decreases the model's flexibility which reduces the model's variance; however, unlike a model without the tuning parameter (i.e. linear regression), it only results in a slight increase in bias compared to a least squares model. However, there is a limit to the size of lambda, as too large of a value can increase the bias at too high of a rate so any benefits are lost. Thus, finding the right balance is important to creating an accurate model and should help improve our ability to accurately predict the donation amount.

Next we will use the lasso technique with the 10-fold cross-validation approach to select lambda. Lasso is very similar to ridge-regression except it can actually select variables because its shrinkage penalty will reduce some coefficient estimates to zero, rendering them insignificant in the model. This can produce a simpler and more interpretable model that is often more advantageous for practical use. Both ridge regression and lasso can outperform the other with different data sets so we use both to determine which is best suited for our data set.

The principal components analysis is a technique for reducing the dimensions of a matrix; it is considered that fewer dimensions will suffice to explain most of the variability in the data. Generally speaking, the reduced number of dimensions selected is considered to be the principal components of the data set. If these assumptions are met, fitting a least squares method to the principal components will often produce better results. Principal components analysis will usually perform well when the first few principal components capture most of the variation in the data, but worse when the opposite is true. Finally, it is important to note that the components are selected in an unsupervised manner.

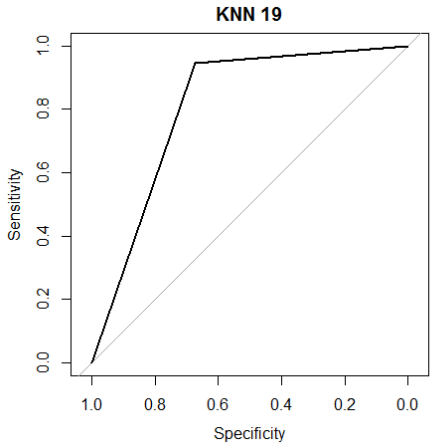
Partial least squares is closely related to principal components analysis with one distinct difference that it uses the response variable to guide the identification of the components. This allows partial least squares to select components that will also be the best for predicting the response; this is a drawback for the use of principal components analysis. In other words, partial least squares tries to approximate the original features well but also tries to relate them to the response variable too.

Results

Classification model results

For our classification model, the goal was to predict the amount of profit and number of mailings required to achieve that profit amount. In order from lowest profit to highest profit, the results from the aforementioned models are listed below.

The model that performed the worse for classification was the K-nearest neighbors (KNN) model. Various KNN models (k = (1, 5, 10, 20, etc.) were tested using different k values to identify the number of neighbors to consider when classifying the response variable (DONR). Ultimately, the best performance, based on maximizing profitability, was achieved when k=19.

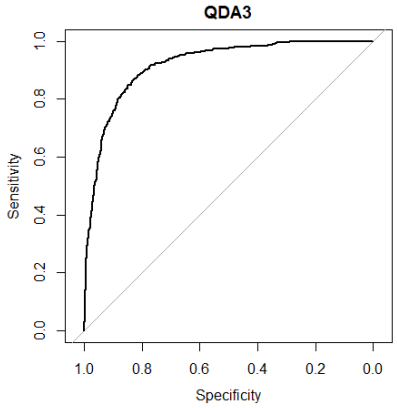


Maximum profit: $11,146.5
Number of mailings: 1278
Error rate: .16501
Sensitivity: .94595
Specificity: .67321
AUC: .8096

Since K-nearest neighbors models often outperform logistic, LDA, and QDA when the boundary line

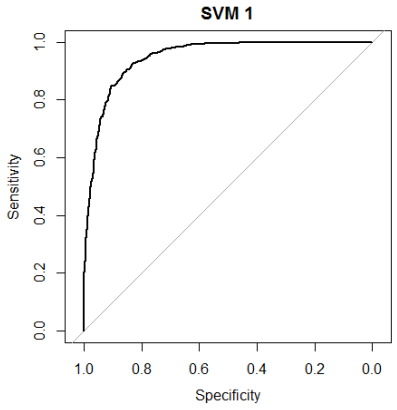
between classes is highly non-linear, this indicates that a more linear boundary line exits in this data set.

The next lowest performing model was QDA. Similar to KNN, this model performs better when there are more non-linear boundaries between classes of the response variable because it assumes a covariance matrix for each variable. Four different QDA models were fit and, interestingly, this was the only model that performed better when some of the variables were removed. Specifically, INCA and LGIF were removed because they showed VIF values higher than 5. Removing these variables only improved the QDA model.



Maximum profit: $11,236
Number of mailings: 1313
Error rate: .19822
Sensitivity: .95696
Specificity: .64966
AUC: .918

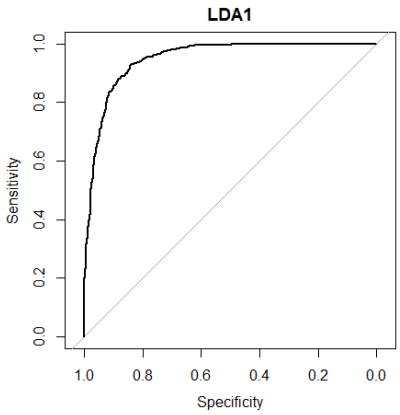
The next model was the support vector machines (SVM). Again, this is another model designed to perform very well on non-linear and complex boundaries. Four different cost settings were fitted (c = (10, 1, .01, .001) using 5000 trees and SVM performed the best when cost=.001. Additionally, the kernel was changed from linear, to polynomial, to radial and the linear kernel performed the best. This is another indication that the boundary is closer to linear and therefore models with more linear assumptions will likely perform better.



Maximum profit: $11,636
Number of mailings: 1374
Error rate: .1893
Sensitivity: .99299
Specificity: .6251
AUC: .9467

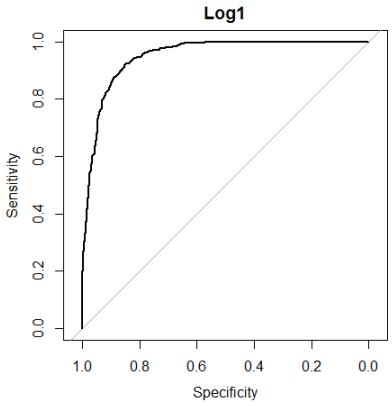
The next highest profit model was the linear discriminant analysis (LDA). We performed another four models with LDA. One only used the subset of variables selected by our

stepwise and best subset selection methods, another removed variables with VIF > 5, and another included polynomial transformations of variables. The model with second degree polynomial variables of TDON and HINC performed the best, and this amount and type of variables will show to perform the better within almost every other type of machine learning model.



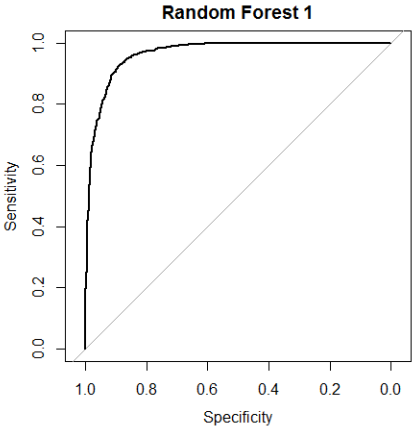
Maximum profit: $11,657.5
Number of mailings: 1356
Error rate: .18483
Sensitivity: .99199
Specificity: .6418
AUC: .9485

The next highest profit model was logistic regression. We fitted three different logistic regression models. One only used the subset of variables selected by our stepwise and best subset selection methods, another included an additional variable, one used the full amount of variables, and another included polynomial transformations of variables. The model with second degree polynomial variables of TDON and HINC again performed the best.



Maximum profit: $11,702
Number of mailings: 1341
Error rate: .17641
Sensitivity: .99299
Specificity: .6575
AUC: .95

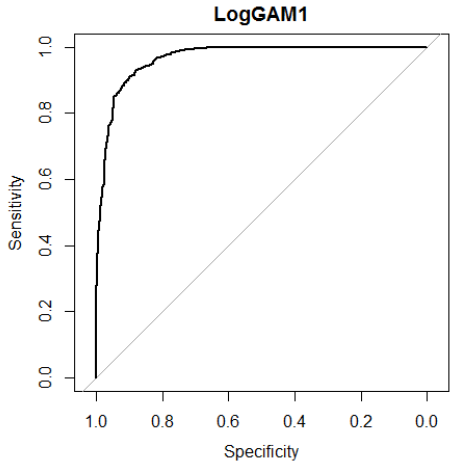
The third best maximum profit was obtained with the random forest model. We fitted many models with different variables and different mtry levels (mtry = (5, 10, 15, 20). Ultimately, we found that using the square root of the total number of variables (22 variables – full model plus two polynomial variables) for mtry (mtry=5) produced the best results. Random forest produced more profit using



over a hundred fewer mailings that logistic regression.

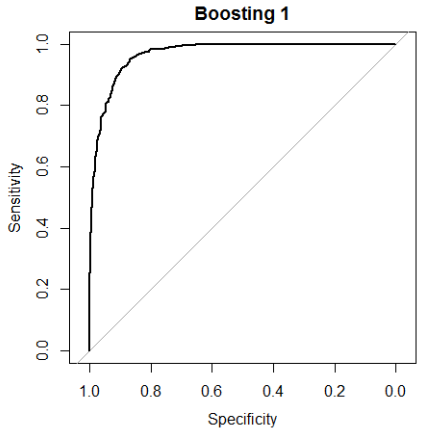
Maximum profit: $11,784.50
Number of mailings: 1220
Error rate: .127354
Sensitivity: .981981
Specificity: .7655
AUC: .965

The second highest profit was found with the general additive model with the logistic regression. We fitted five different models each using various additions of splines to variables that did not appear to have as much of a linear correlation with the response variable. Of the five models, the best model used splines for CHLD, HINC, and TDON. This model performed almost as well as our boosted model, and, with unlimited time to try more variable variations, it's quite possible it could be tweaked to outperform the current boosted model.



Maximum profit: $11,850.50
Number of mailings: 1245
Error rate: .13181
Sensitivity: .98999
Specificity: .74887
AUC: .9663

The highest maximum profit was found by using the boosting approach with a decision tree. Until some extensive tweaking was performed on the GAM model, this model clearly had outperformed the other models in terms of maximizing profitability. Various variable subsets and shrinkage values were fitted with the best model using a shrinkage value of .001 on the full model with one additional second degree polynomial variable of the HINC variable.

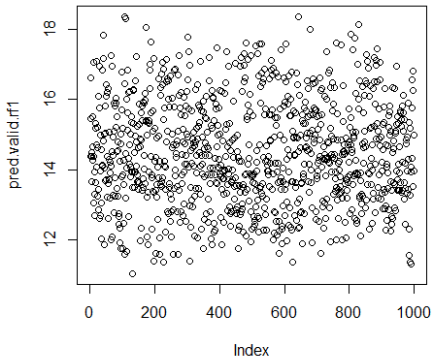


Maximum profit: $11,861
Number of mailings: 1189
Error rate: .1021
Sensitivity: .983
Specificity: .7978
AUC: .9672

Regression model results

For our regression model, the goal was to predict the amount of the donation from a donor. The performance of the model was based on the mean prediction error comparing the predicted amount versus the donation value (DAMT) in the validation data set. In order from lowest profit to highest profit, the results from the aforementioned models from the analysis section are listed below.

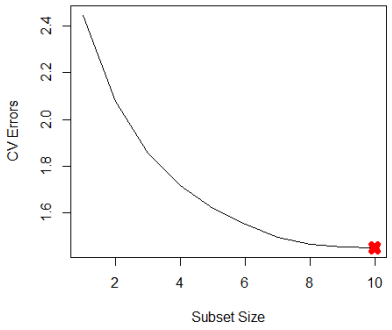
The model that showed the worst mean prediction error was found with the random forest model. Instead of using the square root of the number of predictors as our mtry value we used the number of predictors divided by three for mtry. Unfotuantely, even with different mtry values this model did not perform very well in this context. Using the importance() function, we were able to see that the variables LGIF, RGIF, and AGIF were the most important variables in the random forest model.




```
lgif      33.19328258    1213.89655
rgif      33.80424528    1227.28495
agif      28.77718570    1031.15007
```

Mean prediction error: 1.661631
Standard Error: .1732228

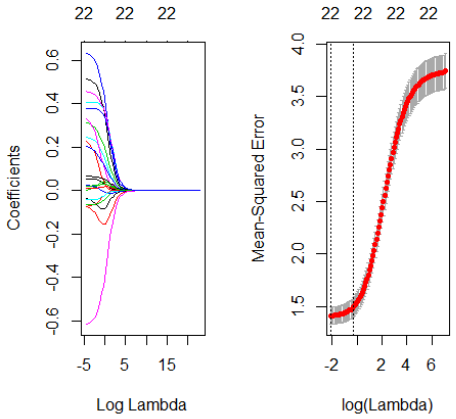
The model that showed the second worst mean prediction error was found using the best subset selection model with a 10-fold cross-validation method. The cross validation errors were the lowest when using a 10 variable model. Since our data set has been standardized for the predictor values, we can compare the relative strength of each variable to the response value, DAMT. Here the coefficients show that region four (REG4) has the largest relative effect on the DAMT.



Mean prediction error: 1.608761
Standard Error: .1643959

```
(Intercept)         reg3         reg4         chld         hinc         incm
 14.1393532   0.3703004   0.7447163  -0.5851488   0.5330191   0.3496308
plow         tgif         lgif         rgif         agif
0.2450706  0.2090670   0.4487916   0.4207344   0.3988884
```

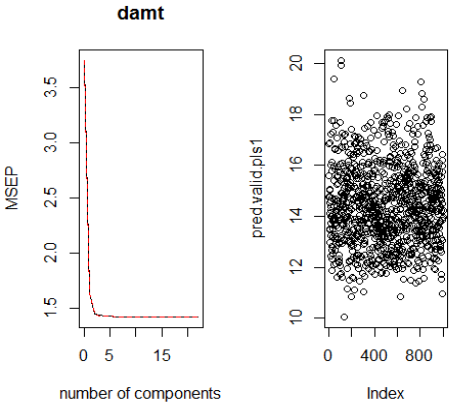
The model that showed the third worst mean prediction error was found with the ridge regression model. Here we used cross-validation to determine the best lambda value to improve our prediction accuracy. The best lambda value was 0.125775. We also experimented between using the full model and one with additional polynomial variables added to the model; the latter model, as was often the case in this project, produced the best performance. The variable with the strongest effect on DAMT was CHLD and had a negative effect on the DAMT value.



Mean prediction error: 1.601929
Standard Error: .1624482

```
chld         -0.56911519
```

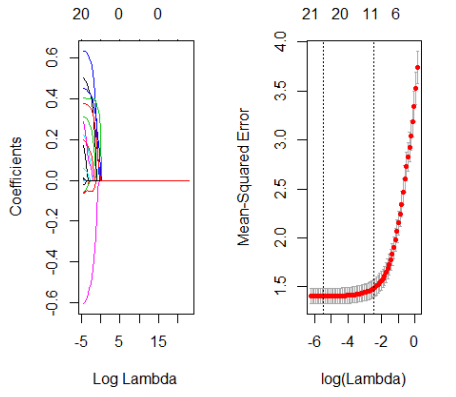
Partial least squares model produced the next best mean prediction error. Again we used cross-validation to determine that seven components had the smallest adjusted CV value. We applied the model to the validation set using 7 components and found the results showed an improvement over the ridge regression model. Consistent with other models, a review of the standardized coefficients shows REG4, CHLD, and RGIF having the strongest effect.



Mean prediction error: 1.593234
Standard Error: .1614317

```
reg4         0.5399287331
chld        -0.5080942644
rgif         0.4612949711
```

Lasso produced the next best mean prediction error. Again we used cross-validation to determine that seven components had the smallest adjusted CV value by determining the best (min) lambda to use. The best lambda value was 0.125775. Here we used cross-validation to determine the best lambda value to improve our prediction accuracy. Consistent with other models, a review of the standardized



coefficients showed REG4, CHLD, and RGIF having the strongest effect.

Mean prediction error: 1.601929
Standard Error: .1624482

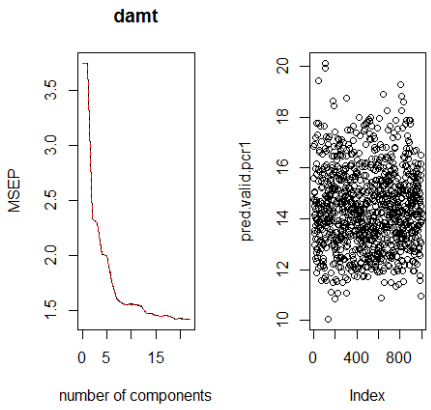
```
reg4       0.636613844
chld      -0.613173694
hinc         0.510028028
```

Least squares produced the third best mean prediction error, proving its robust nature against many formidable models. We used three different models: a full model, one with two additional polynomial variables, and a model with 14 variables. Again, the model with the additional polynomial variables showed the lowest mean prediction error. Consistent with other models, a review of the standardized coefficients showed REG4, CHLD, and HINC having the strongest effect.

Mean prediction error: 1.591109
Standard Error: .11610636

```
  (Intercept)          reg1          reg2          reg3          reg4
14.1834244585  -0.0331643670 -0.0688097903  0.3179403770  0.6387174651
home                  chld          hinc     I(hinc^2)          genf
0.2474793104  -0.6200286358  0.5171875091 -0.0742311970 -0.0663540252
wrat          avhv                  incm          inca          plow
0.0256221483  -0.0415481743  0.3424098971  0.0513823001  0.2363859880
npro                  tgif          lgif          rgif          tdon
 -0.0006251478  0.2066333161  0.4067273793  0.4582684567  0.0661277250
I(tdon^2)          tlag          agif
0.0082538751   0.0212809271  0.3806982622
```

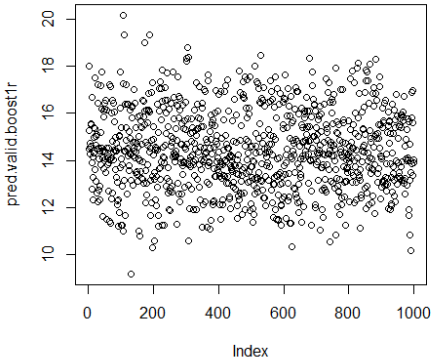
Principal components regression produced the second best mean prediction error for predicting the response variable, DAMT. Again we used cross-validation to determine that the lowest adjusted CV was at 19 comps and again at 21 and 22 components. Since we had identical values for three components we tested each by varying the ncomp value and found ncomp=22 had the best performance. Still consistent with other models, a review of the standardized coefficients showed REG4, CHLD, and RGIF having the strongest effect.



Mean prediction error: 1.591109
Standard Error: .1610636

```
reg4       0.5399287331
chld      -0.5080942644
rgif       0.4612949711
```

Finally, the best mean prediction error was found using the boosting method from regression decision trees. Various shrinkage values were tested (.01, .1, 1, 10) using 5000 trees and .01 produced the best results using 5000 trees. We also tested the full model with the same two additional polynomial variables and a model with a subset of variables; the larger model performed better. Furthermore, we used an interaction depth of 4 and set the distribution value to Gaussian. The power of the boosted method was really apparent as it had a significantly lower mean prediction error than any of the other models.



Mean prediction error: 1.601929
Standard Error: .1624482

```
reg4        0.636613844
chld       -0.613173694
hinc          0.510028028
```

Conclusion

Our recommendation for the charity organization would be to use the boosting method applied to classification and regression trees. In the classification setting, it produced the maximum profitability also while having the lowest number of mailings. However, the general additive model was not far behind in terms of the maximum profitability but did require substantially more mailings, which means more labor for the organization. Yet, if interpretation was to ever become an important issue, while it doesn't produce the maximum profitability, using the general additive model with logistic regression could be a decent substitute for the boosting method on classification trees in this data set.

In the regression setting, there was no question that the boosting method on regression trees produced the lowest prediction error. Thus, the organization should use that model to predict the donation amounts. Also, it should be noted that four variables (REG4, CHLD, HINC, and AGIF) consistently showed as having the strongest standardized coefficients over all the regression models. This information on the variables could be used to direct future or deeper analysis on this data set.

Our hope is that, by conducting this thorough analysis and project, that these models will allow the charity organization to improve their ability to collect donations (maximizing profitability), to do so with less labor (creating and sending out less mailers), and to ultimately help improve their ability to provide aid, assistance, and funding to those in need.

Appendix

Hastie, T., James, G., Tibshirani, R., Witten, D. (2013). *An Introduction to Statistical Learning with Applications in R*. New York, NY: Springer.