Hillary Mahoney
PREDICT 422 – Sec 58
Charity – Course Project

INTRODUCTION

A charitable organization wishes to improve the cost-effectiveness of their direct marketing campaigns to previous donors. According to their recent mailing records, the typical overall response rate is 10%, meaning 1 out of every 10 people respond. Consequently, mailing everyone is not profitable, because the average donation for those who do respond does not cover the cost of the mailings. In order to maximize net profit, the goal is to be able to determine whether an individual will donate or not, and then estimate the amount of the donation for those that are expected to donate.

This was achieved by using their mailing list of 8009 individuals which contained information about each of the potential donors' demographic information, income, geographic location, and past donation behavior. First, exploratory analysis was performed on the individuals in the mailing list to gain general insights and understanding of the data at hand. Then, several classification models were constructed to determine whether an individual was expected to donate or not. In other words, individuals were classified as donors or non-donors. The methods used for these predictions included logistic regression, linear discriminant analysis, quadratic discriminant analysis, random forests, boosting, support vector machines, neural nets, and k-nearest neighbors (KNN). The method that produced the most accurate predictive results on a validation data set and maximized profit was used to determine the expected number of donors, and then separate techniques were used to predict the expected amount each would donate. The methods used for these quantitative predictions included multiple linear regression, partial least squares regression, random forests, neural nets, bagging, and boosting. Similar to the classification step, these models were compared to each other using a validation data set to find the method with the lowest prediction error.

ANALYSIS PLAN

The 8009 observations were split into three subsets, a training set containing 3984 observations, a validation set containing 2018 observations, and a test set containing 2007 observations. Because weighted sampling was used, the training and validation sets contained approximately equal numbers of donors and non-donors due to over-representation of responders. Therefore, the validation set had a mailing response rate of 50%, whereas the test set had the typical 10% response rate that was noted in the Introduction. All exploratory analysis and models were fit to the training data set, and all variables were standardized for the analysis.

First, classification models were created to predict whether an individual was expected to donate or not and the results from these models were compared against the actual classifications available in the validation data set. Model accuracy was assessed using confusion matrices on the validation set which provided the number of correctly classified and misclassified observations. That is, by using the validation set we were able to determine how many observations the various models correctly classified as donors and how many observations were misclassified (predicted to donate, but did not actually donate). Then, based on the observations that were predicted to donate, the number of mailings necessary to maximize profit was calculated and the expected profit was calculated using the average donation amount of $14.50. Models with the highest maximum profit were compared to determine which was the best choice for predicting donor responses.

Second, models were created to predict the donation amount given the observation was expected, or predicted, to donate. The parameter estimates resulting from the models fit to the training data set were used to predict the donation amount in the validation data set, and then these predicted donation amounts were compared to the actual donation amounts. The differences between the predicted and actual observations were squared and then averaged to obtain the mean squared prediction error (MSE), and then used to calculate the standard error of the mean squared prediction error. The mean prediction error (MSE) was defined as the mean of the squared differences between the actual observations and their predicted value in the test data, e.g. the

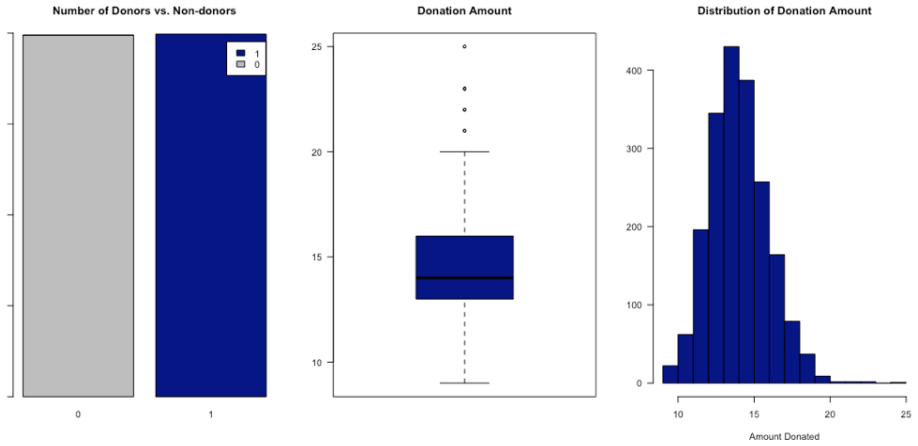
data not used during fitting. The standard error of the MSE was computed by taking the standard deviation of the squared differences between the actual and predicted values in the test data, and dividing it by the square root of the number of observations in the test data. Models with the lowest MSE were compared to determine which was the best choice for predicting the expected donation amount.

EXPLORATORY DATA ANALYSIS

As mentioned, the data was split into three subsets: training set, validation set, and test set. Exploratory analysis was performed on the training set which contained 3984 observations and there were 23 original variables, but only 22 were used for analysis as one was the donor ID number. Variables were examined in regards to their relationship with donor activity to determine if certain characteristics or aspects about the potential donors could indicate whether an individual was going to donate, and if so, the amount of money they would be willing to donate.

Since the goal is to determine whether an individual will donate or not, and then how much they will donate, it makes sense to explore the donor and donation amount variables first. If an observation has donated in the past, they are classified as donor=1 and if an observation has not donated in the past, they are classified as donor=0. Figure 1 shows the number of observations by donor type for the training set, a boxplot of the donation amount from donors, and a histogram of the donation amount from donors. There are almost an equal number of donors vs. non-donors, 1995 and 1989 observations respectively. The boxplot and histogram show that donation gifts above $20 are rare, and those that donate more than $20 are considered outliers. Besides those outliers, the distribution for donation amount appears to be approximately normal with a median donation amount of $14.00 and an average of $14.50.

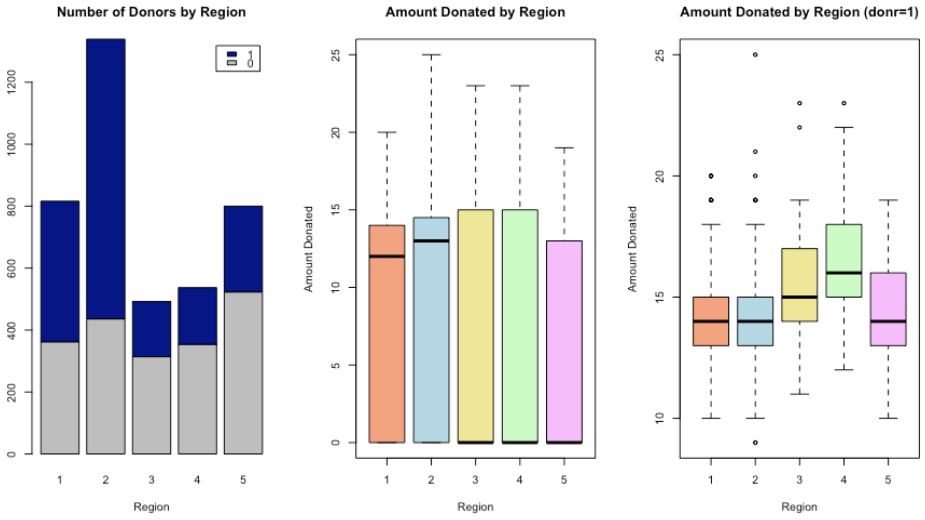
Figure 1: Donors vs. Non-donors and Donation Amount



Starting with region, there are five geographic regions where the potential donor can reside. Figure 2 displays the number of individuals by region broken down by donor or non-donor, a boxplot of the amount donated by all observations by region, and a second boxplot of the donation amount by donors only by region. Region 2 contains the most observations, approximately 34% of the total observations in the training set, and it also contains the most donors (about 45% of the total donors). Looking at the first boxplot that contains the donation amount for all observations by region, there is not a substantial difference in the range of donation amounts by region. However, in the second boxplot that only looks at the donation amount for donors, there is a more distinguishable difference between the regions. Even though region 2 contained the most observations and donors, region 4 has a higher average and median donation amount ($16.44 and $16.00, respectively),

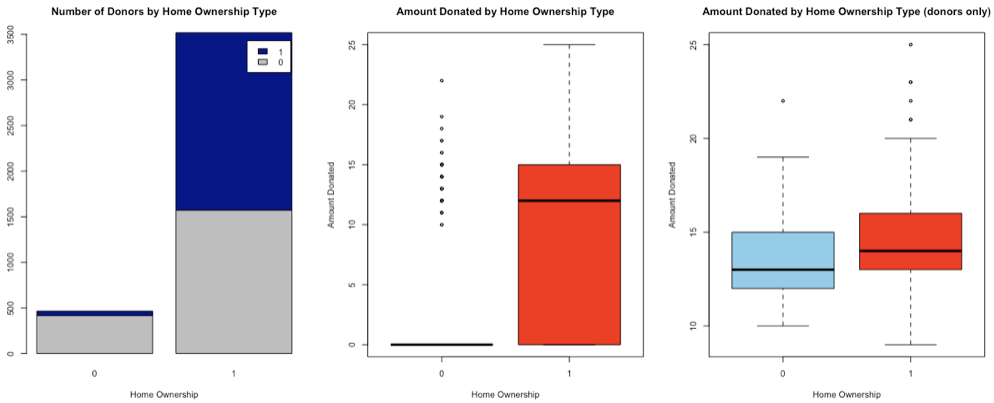
meaning donors in region 4 tend to donate larger amounts than donors in region 2. Additionally, even though region 1 contains almost half the number of donors as region 2, their distributions are very similar, meaning the donation amounts given in region 1 and region 2 are comparable. Lastly, the second boxplot (donors only) also highlights the abnormally large and the one abnormally small donation amounts for each region, which is not visible in the boxplot containing all observations. Given that region 2 contains the most donors, it makes sense there would be greater variability in the donation amounts.

Figure 2: Region



Next, the variable homeowner specifies whether the individual is a homeowner, or not (e.g. a renter). If homeowner=1, then the observation is a homeowner and if homeowner=0, then the observation is not a homeowner. Figure 3 displays the number of observations by homeowner type broken down by donor or non-donor, a boxplot of the donation amount by homeowner type for all observations, and a second boxplot of the donation amount for donors only by homeowner type. It is clearly evident there are more homeowners than non-homeowners and it makes sense there are more donors in the homeowner category than in the non-homeowner category as homeowners may be more financially stable or in a better position to donate. From the second boxplot for donors only, it is apparent that homeowners provide a more charitable donation compared to non-homeowners, but there is also a wider range in the donation amount for homeowners.

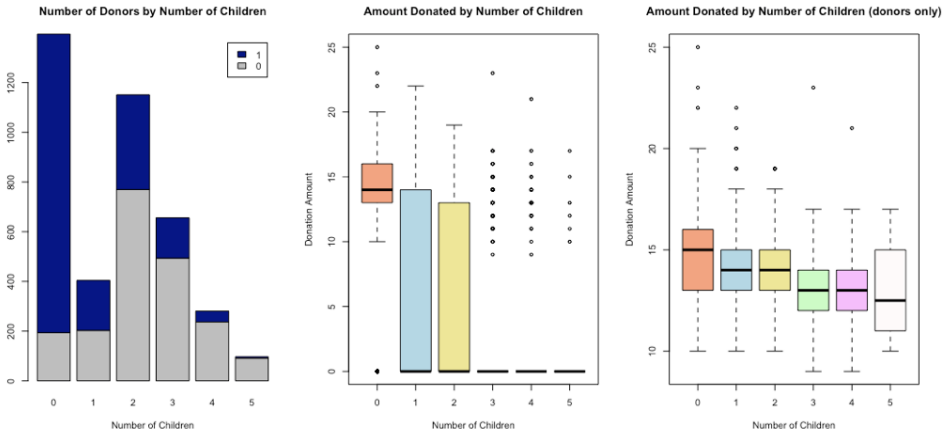
Figure 3: Homeowner Type



The number of children was also assessed in relation to donor behavior. The number of children each observation had ranged from 0-5, with most individuals having zero kids. Figure 4 displays the number of observations by the number of children (0-5) broken down by donor or non-donor, a boxplot of the donation amount by the number of kids for all observations, and a second boxplot of the donation amount by the number of kids for donors only. Approximately 35% of the total observations in the training set did not have any children and about 29% of the observations had two children, these categories also contained the most donors. The likelihood of donation decreased as the number of children increased after two.

The second boxplot shows that donors with one child and donors with two children provide similar donation amounts with the exception of the one child category having more outliers. So even though there are significantly more donors with two children than donors with one child, the donors with two children seem to donate just as much (or as little) as those with one child. We see something similar between donors that have three children and donors with four children; that is, the donation amounts appear to be very similar for observations with three or four children. Lastly, the donors with five children have the widest interquartile range, yet the median donation ($12.50) is very close to the median donation amount from donors with three or four children ($13.00). What this tells us is that even though the number of donors declines significantly for observations with three or more kids, those that are willing to donate typically give similar amounts regardless if they have 3, 4, or 5 children.

Figure 4: Number of Children



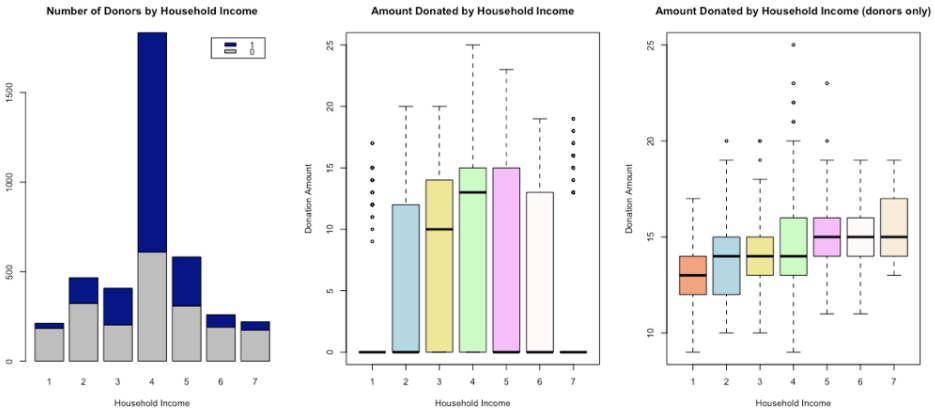
Next, we look at household income which is split into seven categories ranging from 1-7, there is no indication whether category one or seven indicates higher or lower incomes. Similar to prior variables, Figure 5 shows the number of observations for each household income category broken down by donor type, a boxplot of the donation amount by household income for all observations, and a second boxplot of the donation amount by household income for donors only.

The household income category 4 contains just under half (46%) of the total observations in the training set and it contains more than half (61%) of all the donors in the training set. There are not any monetary ranges defined for the income categories, so it's not possible to know the exact household income in dollars for these observations, but the average family incomes for this category is approximately $60,000 and the average median family incomes in the neighborhood is about $47,000. These values were determined by averaging the average family income (inca) for the observations in the category 4 household income, and by averaging the median family income (incm) for the observations in the category 4 household income.

Household income category 4 contains the widest range of donation amounts compared to the other income categories, and it has the most outliers which includes the absolute maximum donation amount of

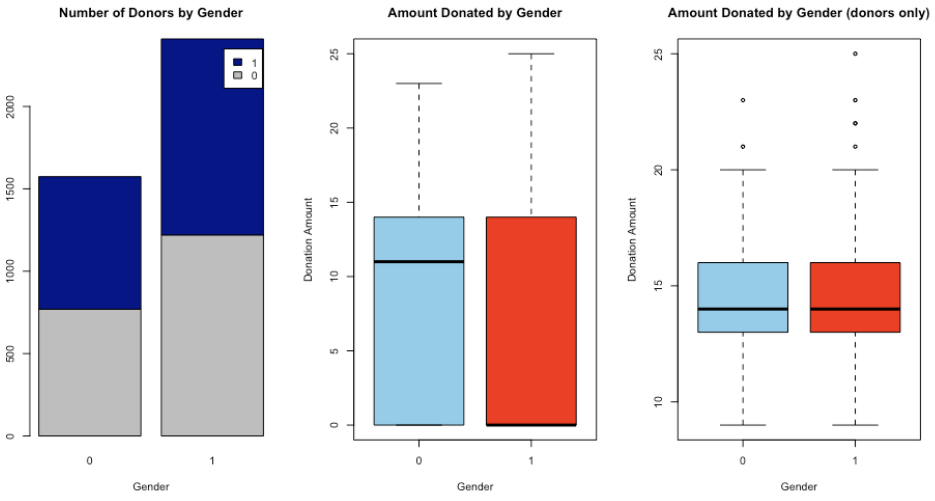
$25.00. However, despite category 4 having the most observations, the most donors, and the widest range of donation amounts, the median donation for category 4 ($14.00) is the same for categories 2 and 3. Similarly, household income categories 5,6, and 7 all have a median donation of $15.00. Even though category 5 contains four times more donors than category 6, they actually have very similar distributions and very similar averages in regards to the amount donated.

Figure 5: Household Income



Gender was also assessed in relation to donor behavior; if an observation is a female they are classified with a 1 and if an observation is a male they are classified with a 0. Out of the total training observations, about 60% are females and 40% are males, but both females and males are almost half and half for donor versus non-donor. That is, out of the females, approximately 49% are donors and out of the males about 51% are donors. Even the donation amounts are very similar for both genders, they share the same median donation amount of $14.00 and the averages are within cents of each other ($14.47 for females and $14.53 for males). This distribution is clearly visible in Figure 6. It seems gender has little to no influence on donor behavior.

Figure 6: Gender

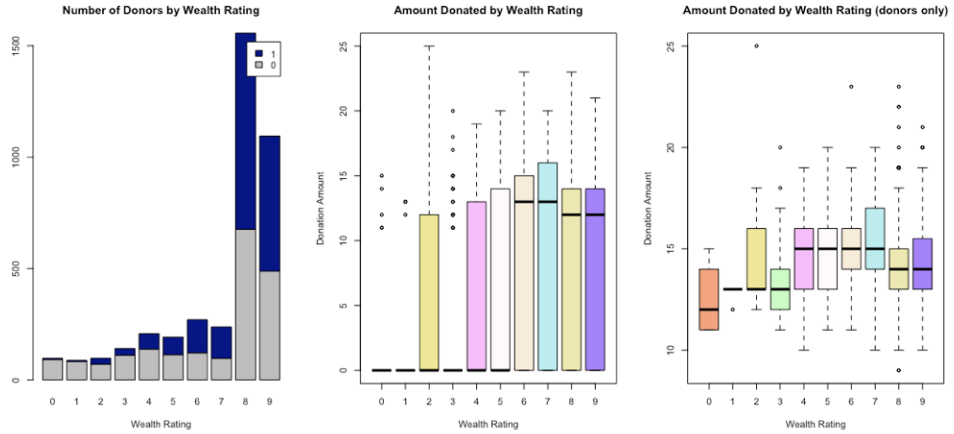


Next, wealth rating was explored. Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with nine being the highest wealth group and zero being the lowest. Figure 7 displays the number of observations for each wealth rating segment by donor type, a boxplot of the amount donated by wealth rating segment, and a second boxplot

of the amount donated by wealth rating segment by donors only. Segments 8 and 9 contain the most observations and the most donors, it makes sense these segments would have the most donors since the individuals who fall in these categories are wealthier thus being more willing to donate. However, it is surprising the drastic and significant difference in number of observations for segments 1-7 compared to segments 8-9. I would have expected more observations to fall in the middle segments rather than the higher wealth rating segments.

Interestingly enough, despite wealth rating segments 8 and 9 containing the most donors and being wealthier, the second boxplot in Figure 7 shows that the donation amounts for segments 8 and 9 are less than the donation amounts from the middle wealth rating segments. Additionally, the likelihood of donating steadily increases from segment 0 to segment 7 (percent of donors goes from 5.7% to 59.2% as wealth rating increases from 0-7), but the likelihood of donating decreases for segments 8 and 9 (drops to 56% and then 55%, respectively). So, it seems wealth rating plays a role in donor behavior but has less of an effect on the amount donated.

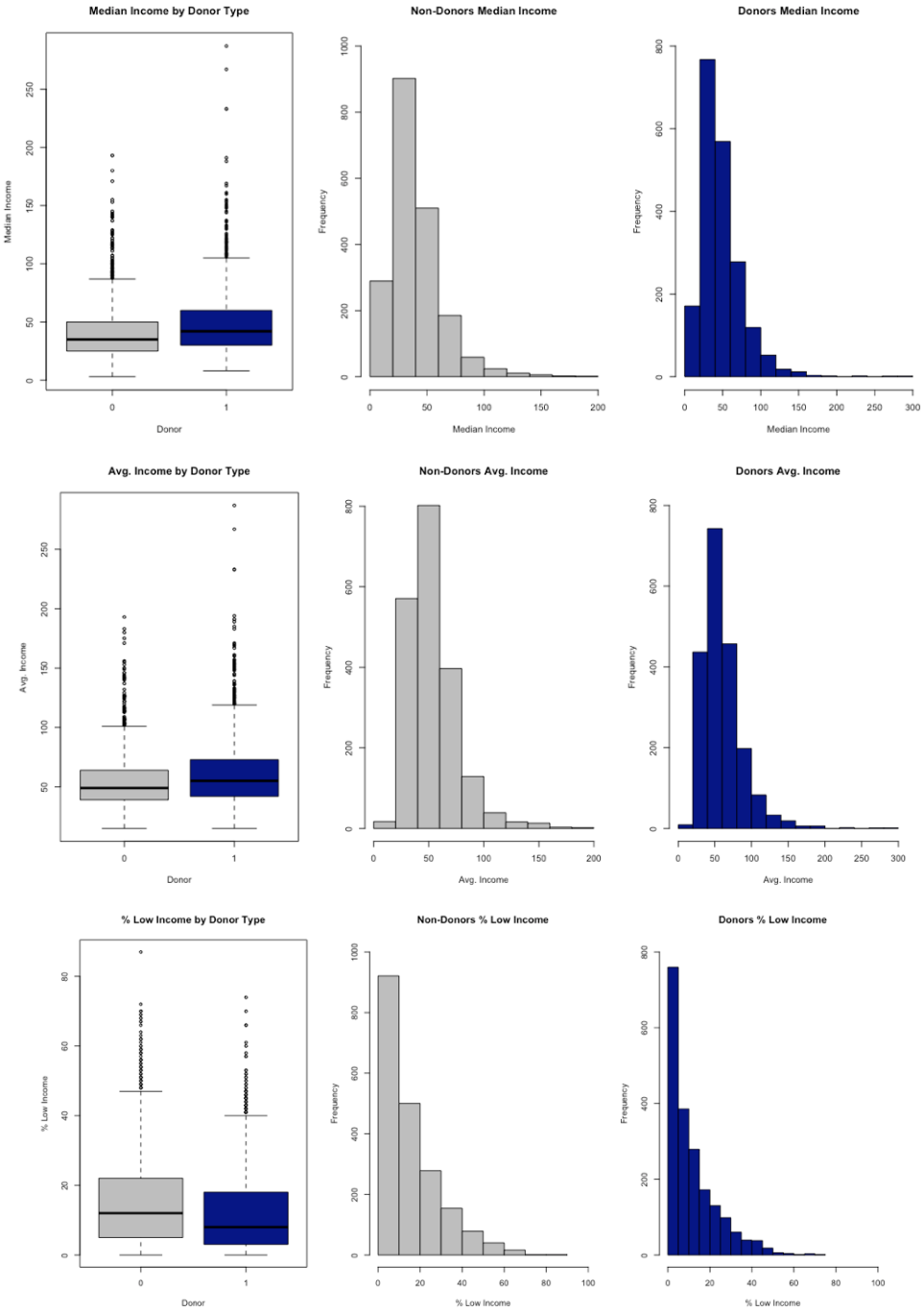
Figure 7: Wealth Rating



Median family income (in thousands), average family income (in thousands), and percent categorized as 'low income' in neighborhood were analyzed in conjunction. Figure 8 first displays median family income in potential donor's neighborhood, then displays average family income in potential donor's neighborhood, and then shows percent categorized as 'low income.' Each variable has a boxplot with percent low income by donor type and then a histogram of percent low income for donors and a histogram for non-donors.

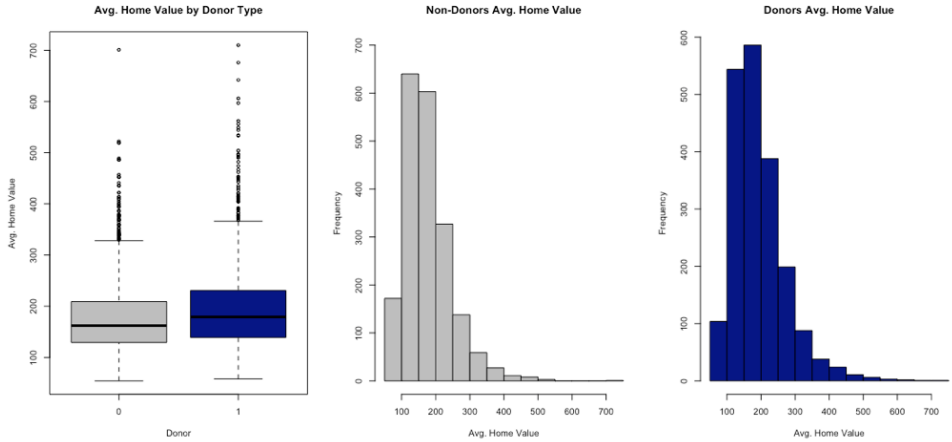
All three variables have quite a few outliers present for both donors and non-donors and all three variables are right-skewed which coincides with the distributions and outliers visible in the boxplots. Donors had slightly higher median and average family incomes than non-donors. For donors, median family incomes that exceed $105,000 were considered outliers and for non-donors, median family incomes that exceed $87,500 were considered outliers. Similarly, for donors, average family incomes that exceed $116,500 were considered outliers and for non-donors, average family incomes that exceed $101,500 were considered outliers. For percent low income, if more than 41% were categorized as low income for donors, then those observations were outliers. Else, if more than 48% were categorized as low income for non-donors, then those observations were considered outliers.

Figure 8: Median Family Income, Average Family Income, and Percent Categorized as 'Low Income'



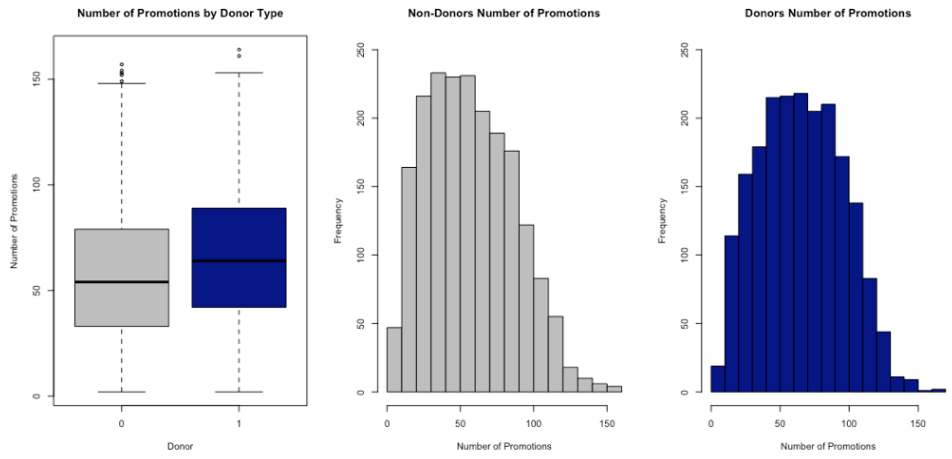
Next, the average home value, in thousands, for a potential donor's neighborhood was evaluated. Figure 9 displays a boxplot of the average home value for donors and non-donors and a histogram for donors and non-donors of the average home value in the neighborhood. Donors had a slightly higher average home value and median average home value than non-donors and quite a few more outliers than non-donors. The distributions of average home value for donors and non-donors were very similar, both right skewed with average home values exceeding $350,000 considered outliers.

Figure 9: Average Home Value in Neighborhood



Lifetime number of promotions received to date is displayed in Figure 10 and was assessed in relation to donor behavior. It is not entirely clear what is meant by promotions, but it was assumed promotions referred to marketing materials. Donors received more promotional materials than non-donors, which makes sense assuming if they had donated in the past then charities they donated to would continue to send them marketing materials. Or, if charities shared donor information with other charities, then it would make past donors more likely to receive additional materials. There is also a strong positive correlation between the number of promotions received and the total dollar amount of lifetime gifts to date. The histogram of number of promotions for non-donors is slightly right-skewed with a few observations having received more than 150 promotional materials in their lifetimes. The distribution for donors is somewhat symmetrical, but it is not the ideal 'bell-shaped' curve that normal or approximately normal distributions have. Similar to non-donors, there were only a handful of observations that received more than 150 promotions.

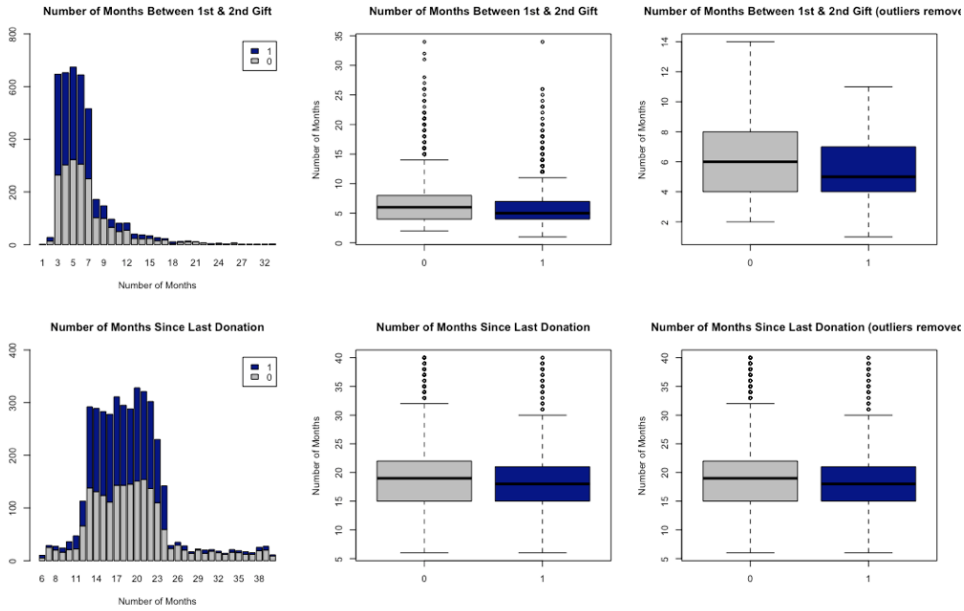
Figure 10: Lifetime Number of Promotions Received to Date



The next two variables deal with the number of months in relation to donation gifts. Figure 11 displays the number of months between the first and second donation gift and the number of months since the last donation. Each variable is displayed in a bar plot by donor type, a boxplot by donor type, and a boxplot with the outliers removed. About 79% of observations donate a second gift within 3-7 months after their first gift and about 84% of observations wait between 12-24 months since their last donation to donate again. There is only a

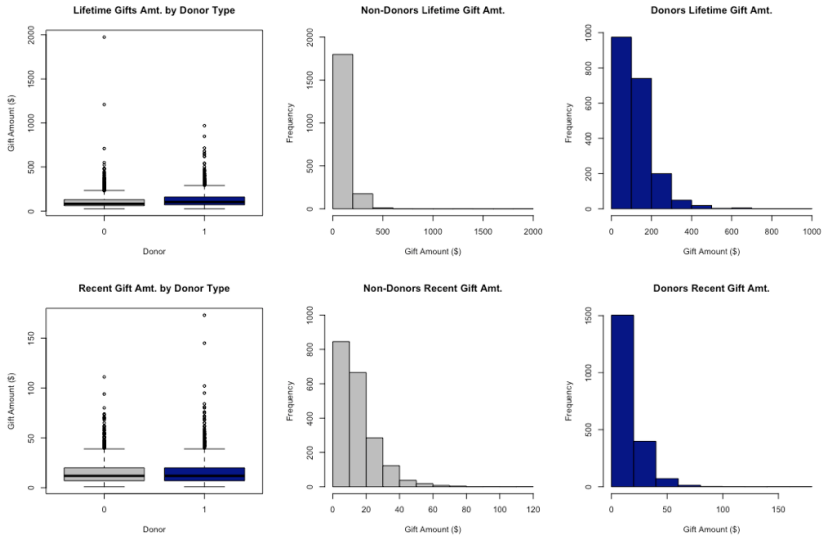
one month difference between the average number of months since last donation for donors and non-donors (18.28 months and 19.35 months, respectively). Similarly, there is only a one month difference between the median number of months since last donation for donors and non-donors (18 months and 19 months, respectively).

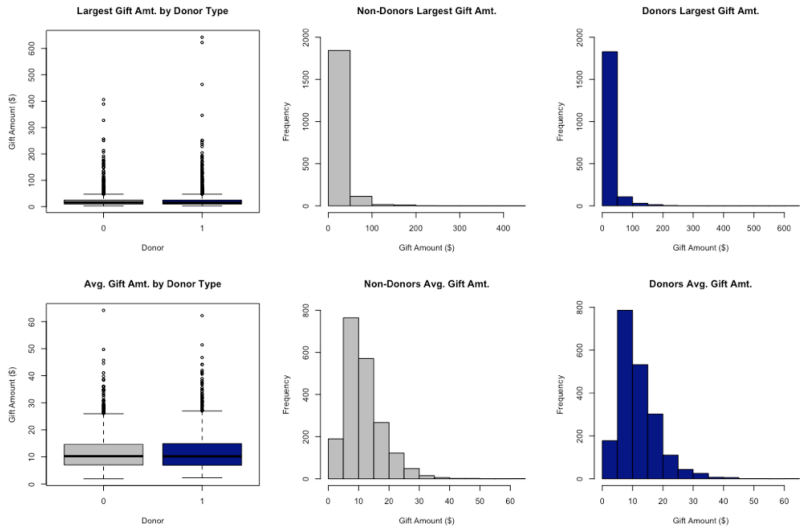
Figure 11: Number of Months Between 1st & 2nd Gift and Number of Months Since Last Donation



Lastly, the dollar amount of lifetime gifts to date, the dollar amount of the most recent gift to date, the dollar amount of the largest gift to date, and the average dollar amount of gifts to date were analyzed together as they all deal with past donation gifts. For each variable, Figure 12 displays a boxplot of dollar amounts by donor type, a histogram of dollar amounts for non-donors, and a histogram of dollar amounts for donors. All of the distributions are skewed right as most of the observations provided donations of $15 or less, very few donated above $25, on average.

Figure 12: Dollar Amount of Lifetime Gifts to Date, Dollar Amount of Most Recent Gift, Dollar Amount of Largest Gift to Date, and the Average Dollar Amount of Gifts to Date



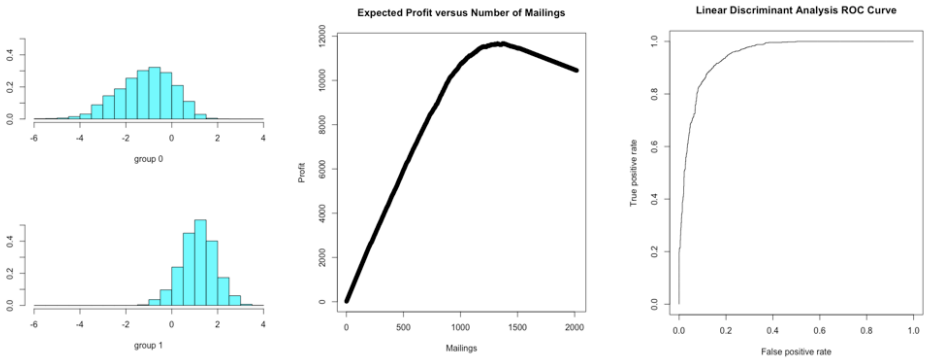


CLASSIFICATION MODELS

Eight techniques were used to construct 19 different models for predicting potential donor response to mailings. All 19 models will not be addressed in detail, but each modeling approach will be discussed along with the best performing model from each technique. First, linear discriminant analysis (LDA) was performed using all original variables (minus the donation amount and the donor ID) without any transformations. From here, adjustments were made, such as variable transformations, to improve the prediction accuracy of the number of mailing responses and to maximize profit. More specifically, the following transformations were used on the predictors: the natural log of average home value (avhv) was used instead of the original average home value variable, household income was squared ($hinc^2$) and included in addition to the original variable, dollar amount of recent gift was cubed ($rgif^3$) and included in addition to the original variable, and lastly, number of months since last donation was squared ($tdon^2$) and included in addition to the original variable.

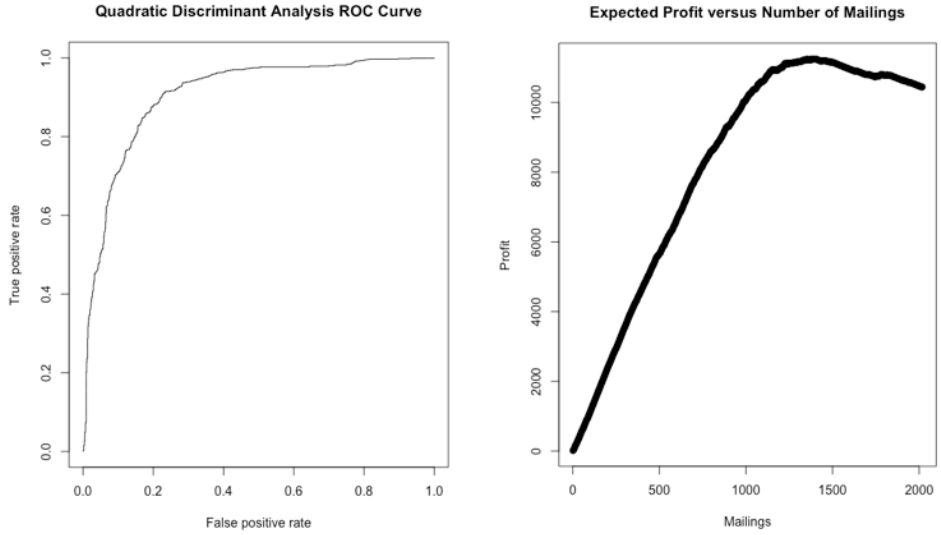
In Figure 13, the histogram shows the distributions for donors (group 1) and non-donors (group 0) and the profit curve shows the expected profit compared to the number of mailings based on a decision boundary of 50% for the posterior probability threshold. The total number of mailings was 1369 in order to maximize profit at $11,675. Out of the 1369 mailings, 994 (72.6%) were correctly predicted as responders and 375 (27.4%) were incorrectly predicted as responders, but were actually non-responders in the validation set.

Figure 13: Linear Discriminant Analysis Output



Quadratic discriminant analysis (QDA) was also used to classify observations on the mailing list. The QDA models performed considerably worse than the linear discriminant analysis models. Similar to the models created using LDA, several approaches were used in order to help maximize profit and increase prediction accuracy. Various combinations of variables and transformed variables were considered, but the best performing model used all of the variables (minus donation amount and donor ID) without any transformations. Again, a decision boundary of 50% was used which resulted in a profit of $11,251 from 1378 mailings to be sent. By using QDA, more mailings were required to be sent, but the expected profit was approximately $424 less than the profit from the LDA model. Only 966 (70%) of the total mailings were correctly predicted to respond, thus 412 (30%) were incorrectly predicted to respond, but did not. Figure 14 displays the expected profit compared to the number of mailings and the ROC curve for the quadratic discriminant analysis model.

Figure 14: Quadratic Discriminant Analysis Output



Next, logistic regression was used to create several models containing different predictors and some transformed predictors. First, all original variables were considered and then the model was paired down to only the variables that were statistically significant. Then additional terms were added by squaring some of the original variables, such as household income ($hinc^2$) and number of months since last donation ($tdon^2$) which resulted in the following equation:

donr = 0.786 + 0.662***reg1** + 1.519***reg2** + 1.364***home** - 2.425***chld** + 0.097***hinc** - 1.105***hinc**^{2} + 0.986***wrat** + 0.460***incm** - 0.283***plow** + 0.558***npro** - 0.563***tlag** - 0.035***tdon** -0.293***tdon**^{2}

The coefficients were then exponentiated and interpreted as odds-ratios, the ratios are displayed in the table in Figure 15 and can be interpreted as follows. If the observation lives in region 1 (reg1), the odds of donating (donr =1) increases by a factor of 1.938. Similarly, if the observation lives in region 2 (reg2), the odds of them being a donor increases by a factor of 4.567. Therefore, observations that reside in region 2 are far more likely to donate than observations that live in region 1. For a one-unit increase in household income (hinc), the odds of an observation being a donor increase by a factor of 1.102. Conversely, a one-unit increase in percent low income (plow) leads to a decrease in the odds by a factor of 0.754. The remaining variables can be interpreted in a similar manner.

Figure 15: Odds-Ratios for Logistic Regression Coefficients

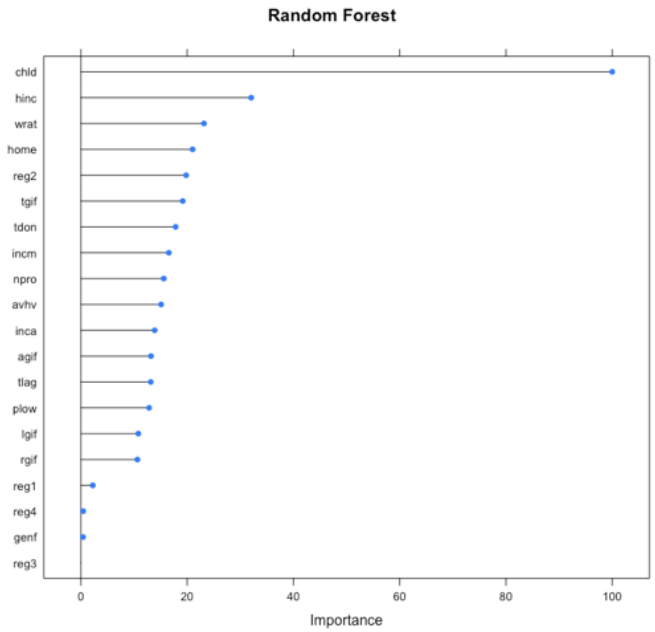
Variable	Odds-Ratio	Variable	Odds-Ratio
(intercept)	**2.195**	**wrat**	**2.680**
reg1	**1.938**	**incm**	**1.584**
reg2	**4.567**	**plow**	**0.754**
home	**3.912**	**npro**	**1.748**
chld	**0.088**	**tlag**	**0.569**
hinc	**1.102**	**tdon**	**0.966**
hinc2	**0.331**	**tdon2**	**0.746**

The logistic regression model produced a maximum profit of $11,689 as a result of 1362 mailings. Both of these are higher than the results from the linear discriminant analysis (LDA) model, but had about the same predictive accuracy. About 73% of the observations were correctly predicted as donors and about 27% were incorrectly predicted as donors, but did not actually donate. So, there was a small increase in profit for no decrease in accuracy between the logistic regression and LDA models.

Next, two decision tree methods were used, first, a random forest was constructed that used 4-fold cross validation and a grid search for tuning. The parameters that lead to the highest average area under the curve (AUC) on the cross folds were selected which lead to a model composed of 500 trees of depth one and three variables available for selection by each tree. The variables are listed and displayed by importance in Figure 16. Variable importance forms a part of prediction power of the random forest model. If the top variable is removed from the model, it's prediction power will greatly reduce. On the other hand, if one of the bottom variables is removed, there might not be much impact on prediction power of the model. This model produced an expected profit of $11,749 from 1245 mailings. Out of the 1245 total mailings, 88.5% of the observations were correctly classified. This is a significant improvement over the other models.

Figure 16: Random Forest Variable Importance

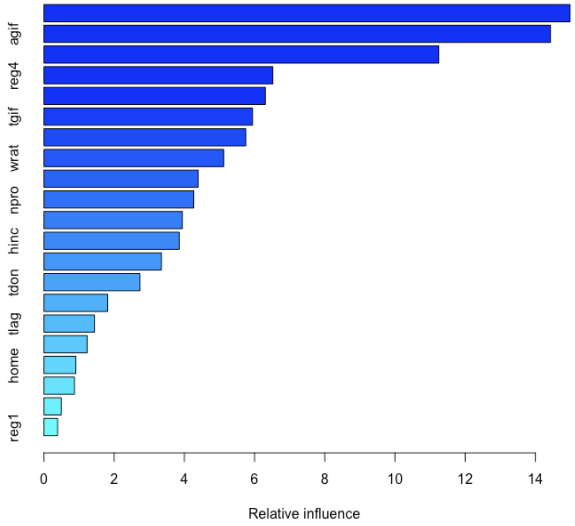
Variable	Overall	Variable	Overall
chld	**100.0000**	**inca**	**13.8837**
hinc	**32.0468**	**agif**	**13.1964**
wrat	**23.1620**	**tlag**	**13.1506**
home	**21.0205**	**plow**	**12.8332**
reg2	**19.8062**	**lgif**	**10.8060**
tgif	**19.1658**	**rgif**	**10.6258**
tdon	**17.8323**	**reg1**	**2.2365**
incm	**16.5399**	**reg4**	**0.4227**
npro	**15.5897**	**genf**	**0.3998**
avhv	**15.0966**	**reg3**	**0.0000**



Second, boosted classification trees were fit to the data to help increase predictive accuracy even more. There were 5000 trees used with a limited depth of four splits for each tree and the variables were hand selected versus using cross validation and tuning over a grid search. The variables included were the same variables used in the linear discriminant analysis model, that is, all variables were included in addition to

household income squared ($hinc^2$), dollar amount of recent gift cubed (rgif3), and number of months since last donation squared ($tdon^2$). Figure 17 displays the variables by their relative influence, dollar amount of recent gift and dollar amount of average gifts to date are the most important variables indicating if they were removed from the model then predictive power would decrease. The boosted model resulted in a profit of $11,874.5 from 1233 mailings and a classification accuracy of 89%. This is an improvement from the random forest and an even greater improvement over the other models.

Figure 17: Boosting Variable Importance



Next, support vector machines (SVM) were used to fit the support vector classifier. A radial kernel was used with a gamma value of 0.02 and a cost of 7 resulting in a wider margin and a larger number of support vectors. Unfortunately, the support vector machine function does not explicitly output the coefficients of the decision boundary obtained when the support vector classifier is fit, nor does it output the width of the margin. The SVM model predicted 1336 mailings with an expected profit of $11,712 and a classification accuracy of 88%. This is less accurate and produces a smaller profit than the boosted and random forest decision tree models, but still better than the linear and quadratic discriminant models and the logistic regression model.

The last five models constructed were three neural networks and two k-nearest neighbors models. The best neural network produced a model with 23 inputs (all of the original variables plus $hinc^2$, $rgif^3$, $tdon^2$), three hidden layers, one output, and 76 weights. This model resulted in an expected profit of $11,388.50 from 1244 mailings. The three neural nets created actually performed the worst out of all the classification models.
The two k-nearest neighbors (KNN) models were created with k=10 and k=12. Both models performed relatively well and were second to the boosted decision tree models in terms of maximized profit, but one of the the downsides to KNN models are they do not provide which predictors are important. The model with k=12 model resulted in the best validation accuracy of 88.5% and predicted 1168 mailings for an expected profit of $11,758.

QUANTITATIVE MODELS

Six techniques were used to construct 10 different models for predicting the amount donated. All 10 models will not be addressed in detail, but each modeling approach will be discussed along with the best performing model from each technique and all model results are summarized in a table in the Conclusions and Recommendations section.

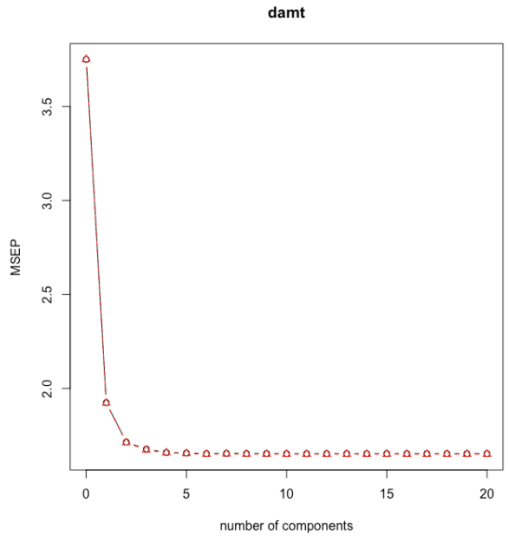
First, an ordinary least squares regression model was fit with all of the predictor variables and was treated as a sort of baseline for improvement. The OLS model was useful to understand which variables were significant, but as expected the model did not perform well. To improve accuracy, non-significant variables were removed. This did improve model diagnostics, but overall the OLS models underperformed compared to the other modeling techniques explored. The adjusted R2 was approximately 0.658 meaning only about 66% of the total variation in the donation amount was explained using the predictors in the model. Additionally, the model accuracy was evaluated using the mean squared error (MSE) on the test set. The MSE is the average squared difference in the predicted versus actual results. Ideally, the lower the test MSE, the better.
The final OLS equation that produced the lowest mean squared error (MSE) of 1.867 is as follows:

damt = 14.159 + 0.362*__reg3__ + 0.671*__reg4__ + 0.248*__home__ - 0.623*__chld__ + 0.500*__hinc__ - 0.064*__genf__ + 0.315*__incm__ + 0.258*__plow__ + 0.185*__npro__ + 0.491*__rgif__ + 0.071*__tdon__ + 0.657*__agif__

The regression coefficients represent the change in the expected donation amount with a one-unit increase for a given variable holding all others constant. For example, if an observation resides in region 3, the expected donation amount increases by 0.362 holding all other variables constant. Similarly, for every additional child, the expected donation amount decreases by 0.632 holding all other variables constant. Lastly, for a one dollar increase in the average dollar amount of lifetime gifts (agif), the expected donation amount increases by 0.657 when holding all other variables constant. The remaining variables can be interpreted the same way.

Next, a partial least squares (PLS) regression model was created with cross-validation using RMSEP as the validation measure. Partial least squares is a supervised dimension reduction method that identifies linear combinations of the original features and then fits a linear model using least squares. The number of combinations, or components, to be used is determined by the RMSEP associated with each number of components. Figure 18 displays the number of components along the x-axis and the RMSEP, which is the cross-validation error, along the y-axis. The graph indicates that the cross-validation error reaches its lowest level when six components are used. Thus, when fitting the model to the validation set, six components were used to predict the donation amounts. This results in a test MSE of 1.868, which is just slightly worse than the OLS model. Additionally, the variance explained from the six components was 57.2%.

Figure 18: Partial Least Squares Components versus Validation Error



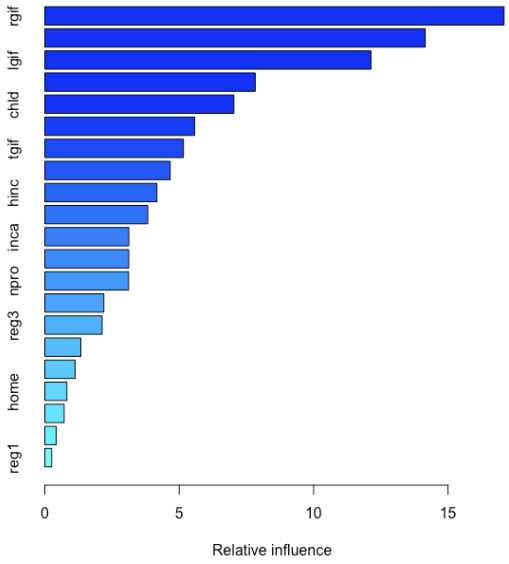
Neural networks were also used in the prediction of the expected donation amount. The neural networks performed better in this setting than when used for classification, but still did not produce the lowest test MSE. Actually, the first neural net which used all of the original predictors plus included household income squared ($hinc^2$) and had two hidden layers produced the highest MSE. However, the second neural net model performed significantly better. It kept the variables the same, but used three hidden layers resulting in a test MSE of 1.604 which is half of what the MSE was for the first neural net.

The last models constructed were all decision trees. The first was a random forest model that was tuned via grid search and 4-fold cross validation. The parameters that lead to the highest root means squared error (RMSE) on the cross folds were selected. This lead to a model composed of 1200 trees of depth 1 and 3 variables available for selection by each tree. This configuration explained about 60.5% of the total variation in donation amount and yielded a MSE of 1.66 on the validation data.

The other two methods used were boosting and bagging which use decision trees as building blocks to construct more powerful predictive models. Bagging helps reduce the variance of a statistical learning method, however bagged trees are typically not an improvement over random forests because random forests decorrelates the trees. So the bagged decision tree produced 500 trees and explained about 60.2% of the total variation and produced a higher MSE (1.705) than the random forest model.

The last model was a boosted decision tree that contained all of the original predictors plus household income squared ($hinc^2$). The shrinkage parameter was set to 0.01 with the depth of each tree limited to four and it was prompted to perform 500 iterations. This resulted in a mean prediction error of 1.46 on the validation data. Similar to the boosted classification tree, the most important variables were the dollar amount of the most recent donation gift and the average dollar amount of total gifts to date. Figure 19 displays the variable importance.

Figure 19: Boosted Decision Tree Variable Importance



CONCLUSIONS AND RECOMMENDATIONS

The model that maximizes the expected profit of the direct mail campaign should be selected from the classification models and the model that minimizes the mean prediction error should be selected from the predictive models, each of the classification models and predictive models are presented in the table below with their respective profit and mean prediction error. Across the eight classification techniques run on the training data, the boosted decision tree performed the best on the validation data, yielding an estimated net profit of

$11,874.50 for a total of 1233 mailings. Similarly, the boosted predictive model performed the best with the lowest mean prediction error, even without using polynomials or interaction terms.

Figure 20: Model Results

Table of final results for classification models:

MODEL	NUMBER OF MAILINGS	PROFIT
LDA1	1329	11624.5
LDA2	1322	11667.5
LDA3	1369	11675
LOG1	1321	11640.5
LOG2	1320	11628
LOG3	1362	11689
QDA1	1399	11238
QDA2	1380	11218
QDA3	1296	11125
RF1	1214	11724
BOOST1	1274	11836
BOOST2	1233	11874.5
SVM1	1336	11712
SVM2	1420	11384.5
SVM3	1323	11665.5
NN1	1311	10773.5
NN2	1125	11017.5
NN3	1244	11388.5
KNN1	1150	11750.5
KNN2	1168	11758

Table of final results for quantitative models:

MODEL	MPE
LS1	1.867523
LS2	1.867433
PLS1	1.868145
RF1	1.672363
RF2	1.656952
NN2	1.604106
BAG2	1.70452
BOOST1	1.539818
BOOST2	1.539335
BOOST3	1.464606

Since the test and validation data sets were oversampled for those donating, an adjustment was applied to the test data to more realistically reflect typical response rates. In fact, the typical response rate to direct mail campaigns is about 10% vs. the 50% observed in the oversampled data. Following the adjustment, the recommended action is to mail 298 individuals with the highest posterior probabilities. Using the average donation amount of $14.50 and the cost of $2.00 per mailing, mailing 298 individuals results in a predicted profit of $3,725.

The analysis conducted was by no means exhaustive, and it is possible that more comprehensive exploration and improvement of models based on other techniques could produce stronger predictions than boosting in the future. Items for further consideration include additional variables, e.g. do the regions contain large cities, encompass multiple states, etc. Having extra insight to the market within each region could prove beneficial to employ different marketing techniques in different regions. Another item for consideration is a study of more systematic ways to evaluate interaction terms and transformations in models (e.g., decision trees for interaction terms), and investigation of techniques for making improvements to decision trees, bagged models, boosted models, and SVMs.