422 Final Course Project

Scott Herman

June 4th, 2017

Prepared for Predict-422: Machine Learning
Northwestern University Masters of Science, Predictive Analytics

Introduction

The purpose of this analysis is to develop a machine learning model to improve the cost effectiveness of direct mailing campaign for a charitable organization. According to their recent mailing records, the typical response rate is 10%, with an average donation of $14.50 from those who respond. The cost to produce and send each mailing is 2.00, which allows us to conclude that it would not be cost-effective for the organizaiton to mail to everyone as this would lead to a negative value for their expected profit. In order to solve this challenge, we will develop two seperate types of models using data from a recent campaign that can effectively capture likely donors so that the expected net profit is maximized. First, we will develop a classification model, which will help us understand which households are mostl likely to respond with a donation. Then, we'll develop a predictive model that will aim to predict the actual dollar amount of the expected donation.

The entire data set consists of 3,984 training observations, 2018 validation observations, and 2007 test observations. Weighted sampling has been used, which over-represents the number of responders so that the training and validation samples have an approximately equal ratio of donors and non-donors. Since we will develop and evaluate our models using the training and validation sets, we will need to account for this inflated response rate when deploying our selected final models on the test set. Our initial data set includes a total of 21 predictor variables which will be used to estimate our two response variables. Our fist response variable, DONR, is binary in nature and indicates whether or not a given household has donated in the past. DAMT, the other response variable we will predict, gives the donation dollar amount received from donor households. The data definitions for each of these variables is given in the table below:

Table 1: Data Definitions

Variable	Description
Reg1	Geographic Region 1
Reg2	Geographic Region 2
Reg3	Geographic Region 3
Reg4	Geographic Region 4
HOME	Homeowner vs. Non-Homeowner
CHLD	Number of Children
HINC	Household Income Category
GENF	Gender
WRAT	Wealth Rating
AVHV	Average Neighborhood Home Value
INCM	Median Neighborhood Family Income
INCA	Average Neighborhood Family Income
PLOW	Percent Neighborhood Categorized as Low Income
NPRO	Lifetime Number of Promotions Received
TGIF	Dollar Amount of Lifetime Gifts
LGIF	Dollar Amount of Larget Gift
RGIF	Dollar Amount of Most Recent Gift
TDON	Number of Months since last donation
TLAG	Number of Months between first and second gift
AGIF	Average dollar amount of gifts to date
DONR	Classifcation Response: Donor vs. Non Donor
DAMT	Prediction Reponse: Donation Dollar Amount

Data Exploration

Our exploratory data analysis begins by examining the structure of our data set. First, we will take a look at the distrubtions of our two target varaiables. Identifying the distributions possessed in the response variables enables us to understand what type of statistical approach will be best suited for yielding the most accurate results in the model development phase. Then, we will plot the distributions of our predictor variables, check for any missing or influential observations, and then examine their correlations to our response variables.

In reviewing the table above along with the histograms below, we can see that DONR is a binary variable, with '1' indicating a donor household and '0' representing a non-donor. This is an important distinction to note as a binary response will not meet the necessary assumptions required for Ordinary Least Squares Regression, and will require a different statistical approach when developing our classification model. On the other hand, we can see that our prediciton response variable, DAMT, is continuous in nature. However, the distribution appears to be zero-inflated indicating the large quantity of households that have donated zero dollars to the charity.

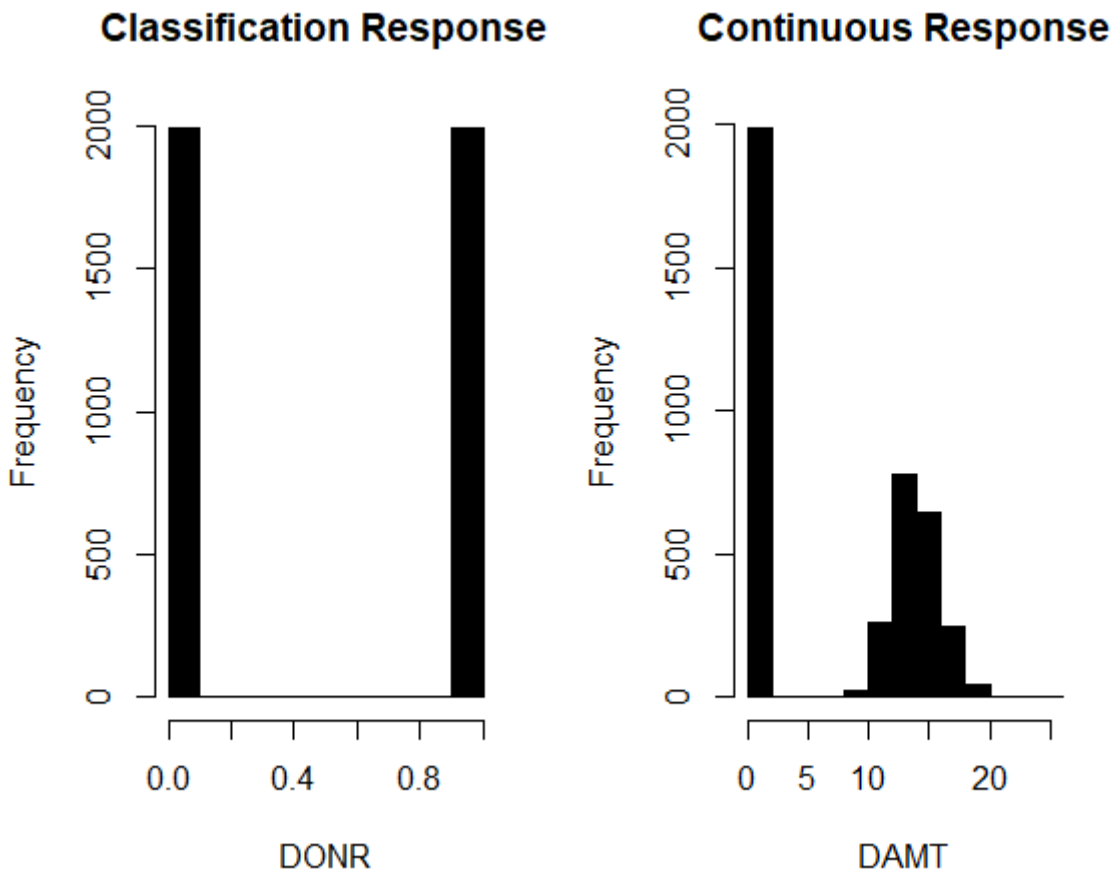

Figure 1: Histogram of Target Variables

Now that we've examined our two response variables, we move on to exploring our set of predictor variables. We note that there are a number of different variable types present within this set, and will proceed first with visualizing our set of categorical variables.

Categorical Variables

The regional breakdown of our sample population is given below in Figure 2. Despite the fact that our data set designates only 4 separate geographic regions, we identified that there are actually five. Those households represented in the fifth region are observations that recorded a zero value in REG1-4. This will allow us to utilize Region 5 as our reference varaible when we reach our model development stage. Here, we can see that REG2, represents the largest popluation of households with nearly 32%, followed by REG5 with roughly 21%, REG1 with 20%, REG3 with 13%, and REG4 with 14%. Additionally, when running a cor.test procedure in R, we found that REG2 showed the strongest positive correlation to DONR with a statistically significant 0.247 correlation value.

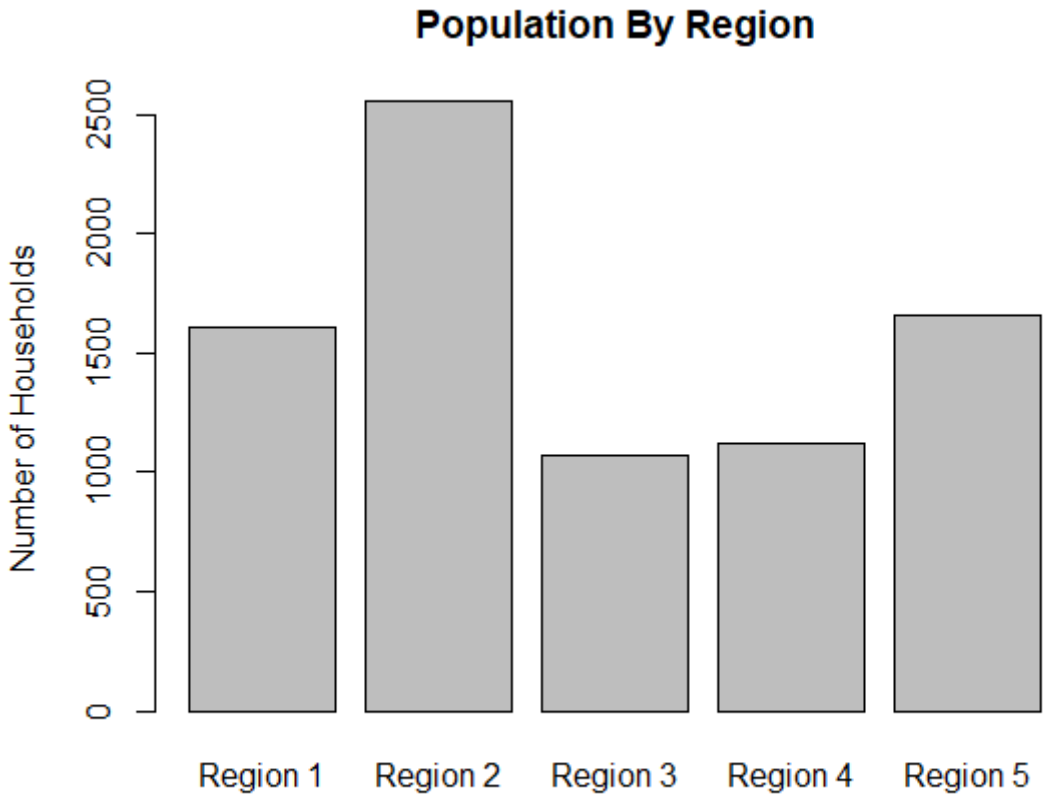

Figure 2: Regional Breakdown

The histograms below illustrate the distriubtions of HOME, CHLD, HINC, and WRAT. Here, we can see that HOME is a binary variable indicating that roughly 13% of our sample population are non-homeowners. Additionally, we can see that the distribution of CHLD tells us that nearly 30% of households do not have children. This could be a significant observation to note, as we also found a significant negative correlation with CHLD and DONR of -0.530 which means that those households that do have children are less likely to donate to the cause. In reviewing the distriubtions of HINC and WRAT, we see that HINC appears to possess a mostly normal distribution, while WRAT appears to be skewed-left with most of the observations falling between the eight and ninth category. We may want to try transforming these variables when we reach the data preparation stage.

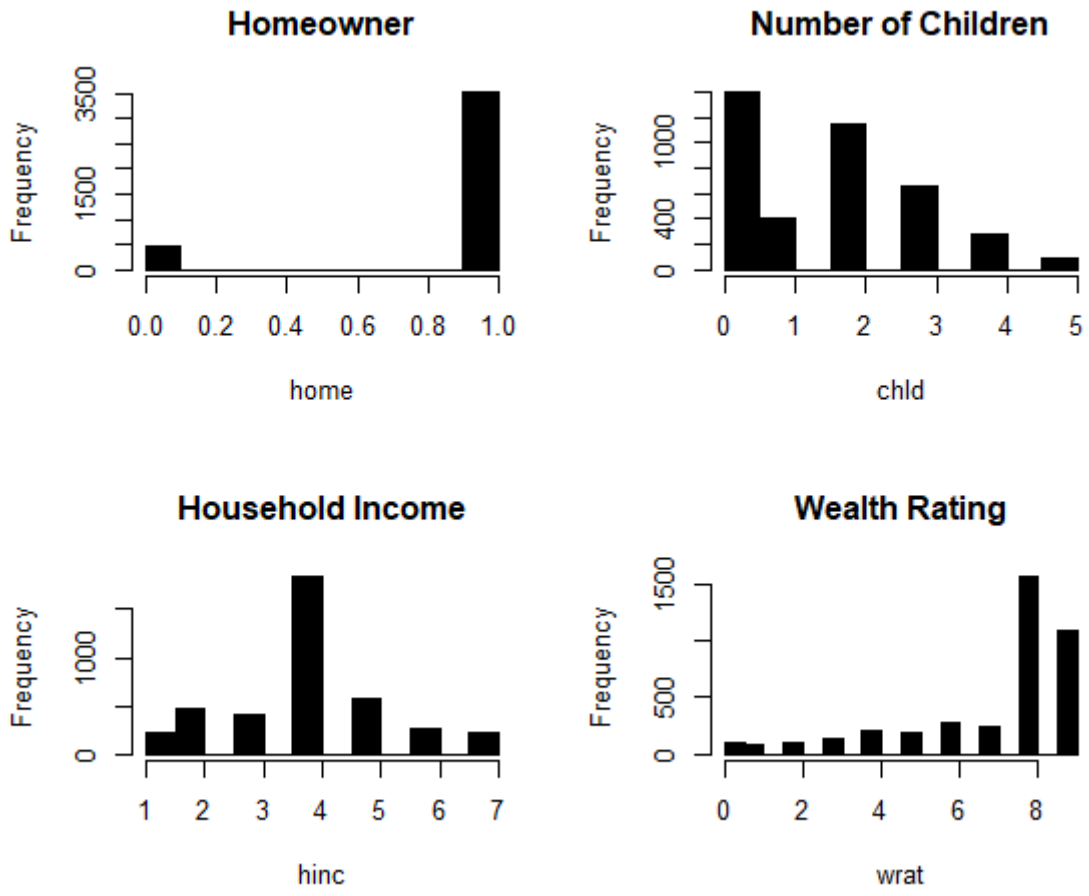

Figure 3: Categorical Variables

Continuous Variables

Next, we explore a set of our continuous variables. In reviewing the scatterplot matrix in Figure 4 below, right away we can see INCM AND INCA possess a nearly perfect linear relationship. This tells us that we may only need to include one of these variables in developing

our models. Additionally, we can see a negative and non-linear association between PLOW and AVHV, INCM, and INCA. This may be another indication of a possible transformation we may want to apply to PLOW in our data preparation stage. We also found positive linear relationships existing between AGIF, RGIF, and LGIF, despite not being shown below.

In addition to understanding the distrubtions of these variables, we also plotted a number of boxplots to understand the potential presence of outlier values that will need to be accounted for. The boxplots of our set of continuous predictors revealed extreme values for AVHV, INCM, INCA, PLOW, NPRO, TGIF, LGIF, RGIF, TDON, TLAG, as well as AGIF. Again, we may want to look at adjusting these outlier points in preparing our data before developing our models.

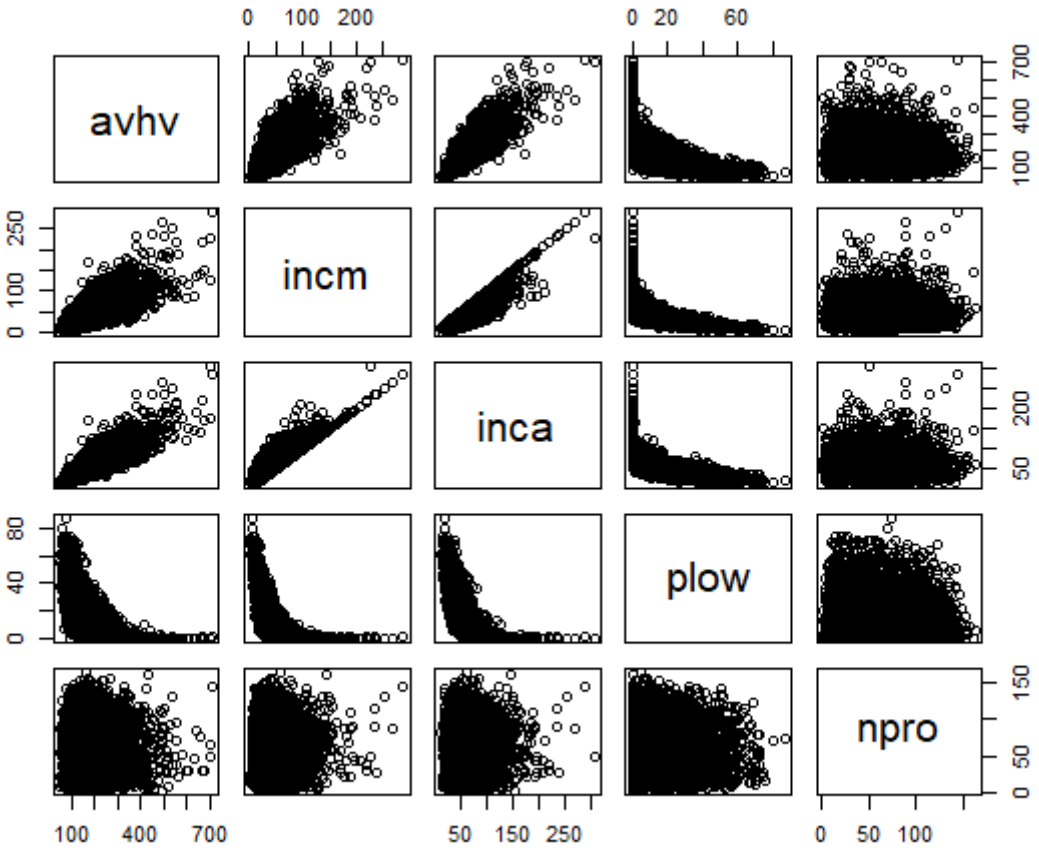

Figure 4: Scatterplot Matrix of Continuous Variables

Data Preparation

In exploring our data set it was determined that our data set contains a number of variables with skewed distributions and influential observations that will need to be addressed, as well as one predictor that we will attempt to re-classify. This step or our modeling process aims to identify how and where to properly adjust these variables in a way that will maintain the consistency of our original data set, while also allowing us to maximize the accuracy in our model results.

First, we have decided to reclassify CHLD, and change this to a binary variable that indicates whether or not a household has children. Given that we saw a strong negative correlation to whether or not a household will donate, we feel this change will help simplify the interpretation and results of our classification model. Reclassifying this variable will result with 30% of these values with a record of '0' for no children, and the remaining 70% representing the precence of children.

Next, we move to variable transformations. We found several variables with skewed distributions which we'll want to transform in order to better meet the assumptions required in our modeling procedures. Although we did identify the presence of outliers, we feel that transforming these variables will be a more appropriate adjustment. The table below highlights the type of transformations we utilized on a selected set of predictor variables.

Table 2: Variable Transformations

Variable	Transformation
AVHV	Log
INCM	Log
INCA	Log
NPRO	Square Root
TGIF	Log
LGIF	Log
RGIF	Log
TDON	Log
TLAG	Log
AGIF	Log

The histograms in Figure 5, below, illustrate the distributions of a number of our transformed set of variables. As you can see, this creates a more normally distributed set of variables which should allow us to avoid some of the struggles seen when attempting to model predictors that possess skewed distributions. Though, we will be able to validate this assumption now that we are ready to move on to developing our models.

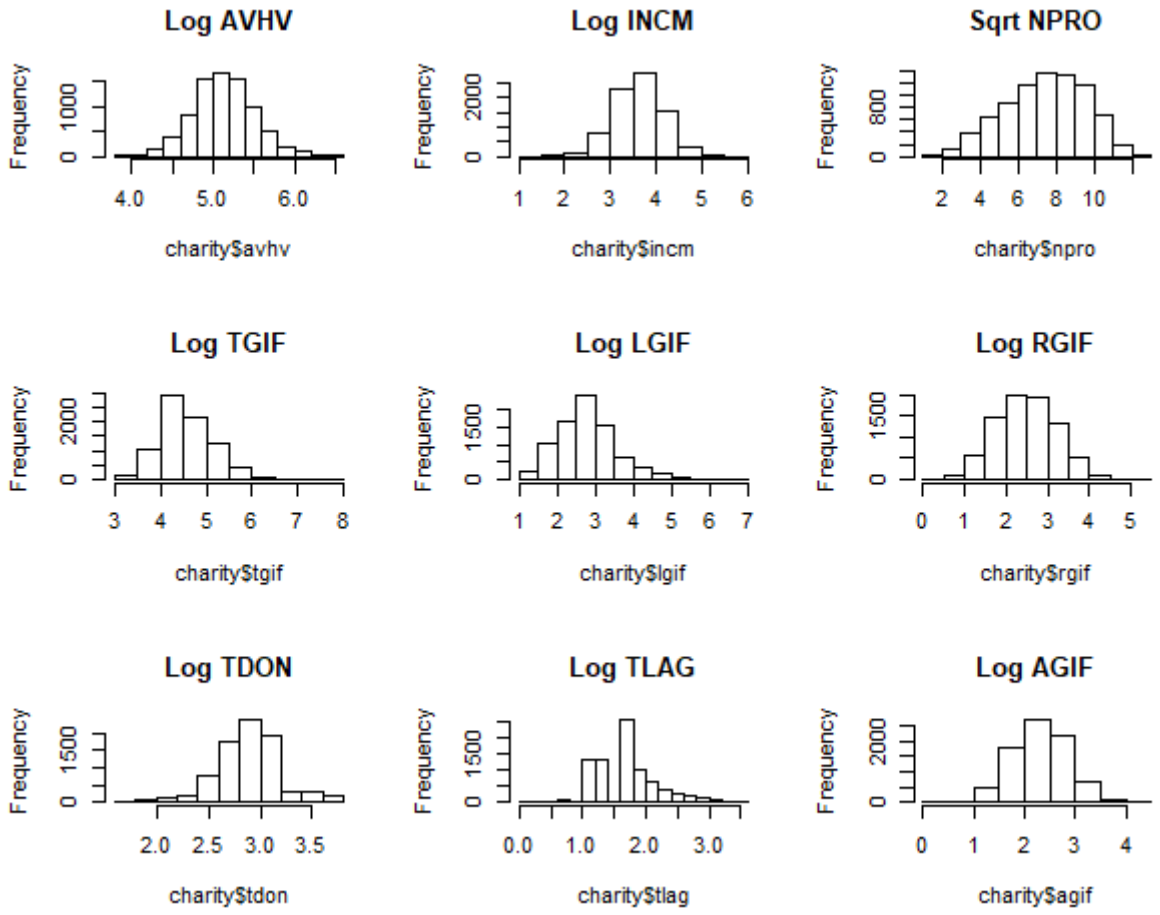

Figure 5: Transformed Variable Distrubtions

Model Development

In proceeding with developing our models, we will attempt to build five separate model types for both our classification and prediction models. For this stage in the process, we created separate training, validation and test sets for each model. We will begin by fitting each model using the training sets, evaluate their results on the validation sets, and then deploy our final predictions on our test set. Additionally, we noted earlier that weighted sampling has been used in our training and validation samples which leaves us with a near equal ratio of donors and non-donors. Therefore, we will need to run a oversampling calculation to account for the true response rate of 10% when we test our final predictions. We will begin first with our classification models, and then move forward to our prediction models.

Classifcation Model Development

In developing our classification models, five separate statistical modeling methods were utilized which allowed us to develop a number of different models for each type. The evaluation criteria in comparing the results of each of this models was based upon the calculated maximum profit predicted. First, we utilized a logistic model using all of the variables within the training set. After reviewing the initial model results, we identified a select set of predictors that proved to be statistically significant. We then reduced the number of predictors in an attempt to understand whether or not this would improve the performance of this model. Once we modified the model, we did in fact find that the model that utilized fourteen statisitcally significant varaibles reduced the AIC score, while also improving the predicted max profit. This model achieved an AIC value of 2060 while predicting a maximum profit of 11670 with 1357 mailings.

Next, we developed a Linear Discriminant Analysis. Although this type of procedure is generally only utilized on qualitative predictors, we tried running this on a both qualitative and quantitative predictors, and began this procedure using the entire set of variables. Again, we attempted to improve the initial model performance by reducing the set of variables and ended up selecting the same set of fourteen variables chosen in our logistic model. The results produced by this model did lead to an improved LDA model which predicted a maximum profit of 11617 with 1340 total mailings. However, these results were not as strong as what we were able to achieve with our logistic model.

Moving on, the next type of model developed was a tree based model. This model was able to correctly classify 85% of the observations, and lead to a maximum profit of 11140.5 with 1165 mailings. Although we attempted to improve the predicted maximum profit by pruning the tree, we found that reducing the number of terminal nodes actually lead to an increase in the error rate while also decreasing our projected profits. In an attempt to improve the performance of our tree model we moved forward with a random forest bag model. Although the bag model did reduce the error rate to 0.1119921, we did not get the profit increases we were hoping for with a reduced profit of 10936.5 with 1035 mailings. Again, we aimed to improve this model's performance through boosting. While the results achieved from the boosted model did improve upon the maximum profit predicted by our bag model, we found

it interesting that this did slightly increase the error rate to 0.1164519. The boosted model led to a predicted maximium profit of 11167 with 1072 mailings. The table below summarizes the results of our classification models

Table 3: Classification Model Results Summary

Model	Number of Mailings	Maximum Profit
Logistic	1357	11670
LDA	1340	11617
Tree	1165	11140.5
Bag	1035	10936.5
Boost	1072	11167

Prior to proceeding with developing our predictive models, we created the final predictions for classifying DONR in the test set. Here, we adjusted for oversampling and made a calculation to account for the true resonse rate which was closer to 10%. In this step, we set a cutoff based upon the number of mailings we acheived in our logistic model, which resulted in a classification vector with 1634 non-donors and 373 potential donors.

Predictive Model Development

In moving forward with developing our prediction models, we will again attempt to build five separate types of models. The evaluation criteria we will use to assess the strength of the prediction models will be based upon the Mean Prediction Error resulting from each model. In this instance, we attempt to minimize the MPE and thus, a lower value indicates a stronger prediction model. The first model was developed through Ordinary Least Squares regression and utilized a total of 20 predictors. We decided to use each of the predictors as the results from this initial model would help serve as a baseline for model accuracy. The OLS model yielded an adjusted R-squared value of 0.621 and a MPE of 1.608413 on the validation set.

Next, we move to building a model using Bet Subset Selection with Cross-Validation. This procedure utilized the regsubsets() function to identify the best subset of predictor variables that aims to maximize the model's accuracy by minimizing the error rate. This procedure found a select set of 12 predictors and resulted in a MPE of 1.601684, which is a slight improvement from our initial OLS model.

Now, we move on with a Random Forest model. This model utilized the full set of predictor variables and explained 59.5% of the variance that occurs in predicting our response of donation amount, which is slightly less than the similar interpretation of the R-squared values found in our initial two models. This slight drop-off in variance explained makes sense logically, as the Random Forest model also led to an increased MPE of 1.713416 which points to a model that performs slightly worse than what we acheived earlier. The next model was

developed using bagging techniques. This model also utilized 20 total predictor variables and led to a MPE of 1.768922, which also did not lead to an improvement.

The last model was developed using boosting techniques, which utilized 5000 trees and led to lowest overall Mean Prediction Error of each of our predictive model. The MPE for this model was 1.573884, which indicates that the boosted model yields the model with the greatest predictive power.

Table 4: Predictive Model Results Summary

Model	Mean Prediction Error
OLS	1.608413
Best Subset with CV	1.601684
Random Forest	1.713416
Bag	1.768922
Boost	1.573884

Results and Conclusion

The results from our logistic regression classifcation model predicted an 18.5% return rate, which is a bit higher than the typical donor response rate of 10%. Additionally, the results from our predictive model utilizing boosting techniques lead to an average donation amount of 14.27, which is slightly lower than then 14.50 average the charity has received in the past. This inidcates that the charity may be able to increase their typical response rate, while accepting a slight drop-off in the average dollar amount received from donor households. Ultimately, this may allow them to boost their overall profitabliltiy with each mailing campaign.