**Clarity in Charity:**

**Classification and Prediction Analysis to Improve**

**A Charity’s Direct Marketing Campaign**

By Ross Walendziak & Sam Neil

## Table of Contents

[Data Exploration](#_edvqfecyackv)

[Introduction](#_wnmo12f12dgl)

[Dependent Variables](#_sb6c9mszo77k)

[Neighborhood Predictors](#_n9j99aggzpen)

[Donor Variables](#_vnpzqigijj5e)

[Prior Gift Variables](#_1c3tjfn4v033)

[General Trends in Predictors](#_syc4peyxqkgo)

[Data Preparation](#_pc5vt8roy2op)

[Classification Models](#_ukyp308vfruv)

[Prediction Models](#_v9fud5oevun8)

[Results & Conclusion](#_pjvvi8twwo8f)

### 

### Data Exploration

#### Introduction

Mailing drives are a vital avenue for charities to generate revenue. The issue with mailing drives is that each mailing costs a fixed amount of money regardless of the applicant’s response. A charity can quickly lose money on mailing drives if they do not focus their campaign on individuals that are more likely to donate. This report will cover various classification models and predictor models to predict if an individual will donate, and if so, how much they would donate. These models are based on a charity’s dataset of 8009 potential donors across 22 variables. This generates 176,198 unique data points. 3,984 of the observations are in the training dataset, 2,018 in the validation dataset, and 2,007 in the test dataset. These variables can be broken up into several broad categories to provide unique insights into the data. It is vital to note that there is no missing data in this dataset besides the test values for DONR and DAMT, which are set to “NA”.

#### Dependent Variables

While there are 22 variables, the first variable is simply the ID number of the potential donor and will not be used in the model. There are two dependent variables for this dataset. DONR will be used as the classification response variable 1 = Donor and 0 = Non-donor. DAMT will be used as the prediction response variable and is the donation amount in dollars. The information for these variables is summarized in the table below:

|  |  |  |
| --- | --- | --- |
|  | DONR | DAMT |
| Minimum | 0 | $0 |
| 1st Quartile | 0 | $0 |
| Median | 0 | $0 |
| Mean | 0.4988 | $7.209 |
| 3rd Quartile | 1 | $14 |
| Max | 1 | $27 |
| Standard Deviation | 0.5 | $7.36 |
| NAs | 2007 | 2007 |

It is important to note that there is no information for these values in the test dataset as they are what the classification and prediction models will predict. The remaining variables will be grouped by their general purpose for ease of reading.

#### Neighborhood Predictors

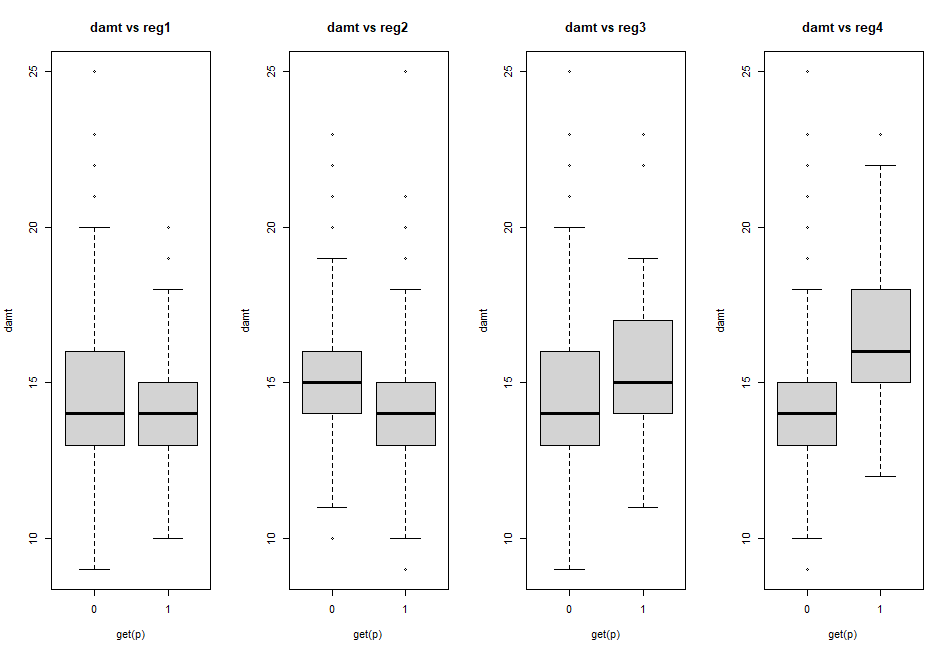
This section will focus on variables related to the neighborhood that the potential donor is from. This will include the following variables:

|  |  |
| --- | --- |
| REG1 | *Region (There are five geographic regions; only four are needed for analysis since if a potential donor falls into none of the four, he or she must be in the other region. The inclusion of all five indicator variables would be redundant and cause some modeling techniques to fail. A “1” indicates the potential donor belongs to this region.)* |
| REG2 |
| REG3 |
| REG4 |
| AVHV | *Average Home Value in potential donor's neighborhood in $ thousands* |
| INCM | *Median Family Income in potential donor's neighborhood in $ thousands* |
| INCA | *Average Family Income in potential donor's neighborhood in $ thousands* |
| PLOW | *Percent categorized as “low income” in potential donor's neighborhood* |

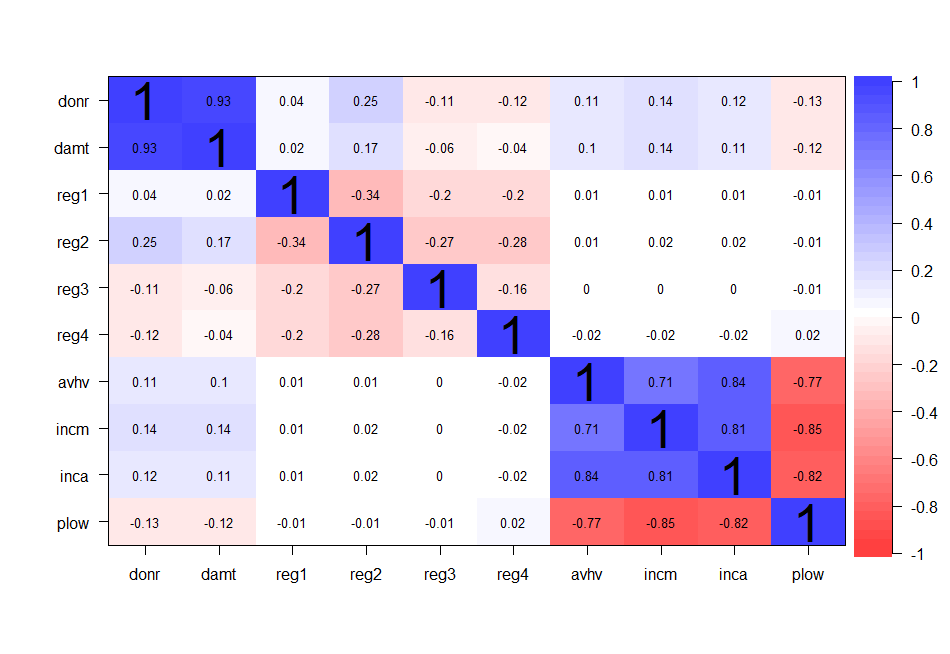
The summaries of these categories for the training data are described in the table below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Min* | *1st Qu* | *Median* | *Mean* | *3rd Qu* | *Max* | *SD* |
| *reg1* | *0.000* | *0* | *0* | *0.2048* | *0* | *1* | *0.4058* |
| *reg2* | *0.000* | *0* | *0* | *0.3361* | *1* | *1* | *0.4720* |
| *reg3* | *0.000* | *0* | *0* | *0.1235* | *0* | *1* | *0.3304* |
| *reg4* | *0.000* | *0* | *0* | *0.1348* | *0* | *1* | *0.3397* |
| *avhv* | *54.000* | *134* | *171* | *185.2* | *219* | *710* | *75.7769* |
| *incm* | *3.000* | *27* | *39* | *44.29* | *55* | *287* | *25.3408* |
| *inca* | *15.000* | *40* | *52* | *57.14* | *68* | *287* | *25.5689* |
| *plow* | *0.000* | *4* | *10* | *13.73* | *20* | *87* | *13.2656* |

The bar charts for the Reg variables are below to highlight the differences in each region.



The correlation matrix for these variables shows that none of the variables are great predictors for donr or damt. There is also high multicollinearity between plow and the other financial measurements of the potential donor’s neighborhood with strong negative correlations. The three other financial measurements of the neighborhood have strong positive correlations up to 0.84 (between inca and avhv). These relationships will be important to keep in mind while building models.



#### Donor Variables

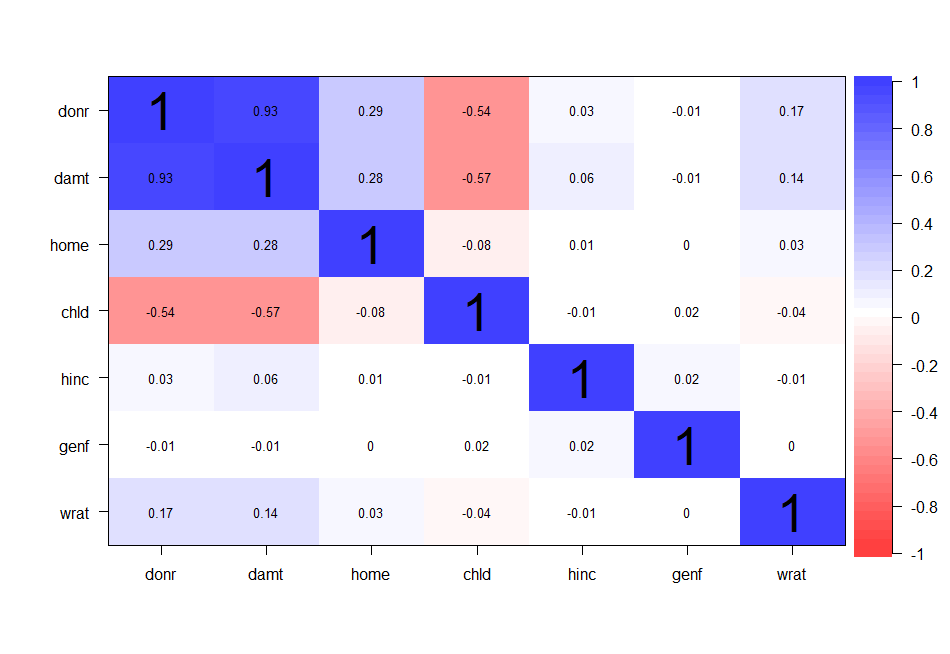
This section will focus on traits related to the potential donor. This will include the following variables:

|  |  |
| --- | --- |
| HOME | *Home ownership status (1 = homeowner, 0 = not a homeowner)* |
| CHLD | *Number of children* |
| HINC | *Household income (7 categories)* |
| GENF | *Gender (0 = Male, 1 = Female)* |
| WRAT | *Wealth Rating (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 9 being the highest wealth group and 0 being the lowest.)* |

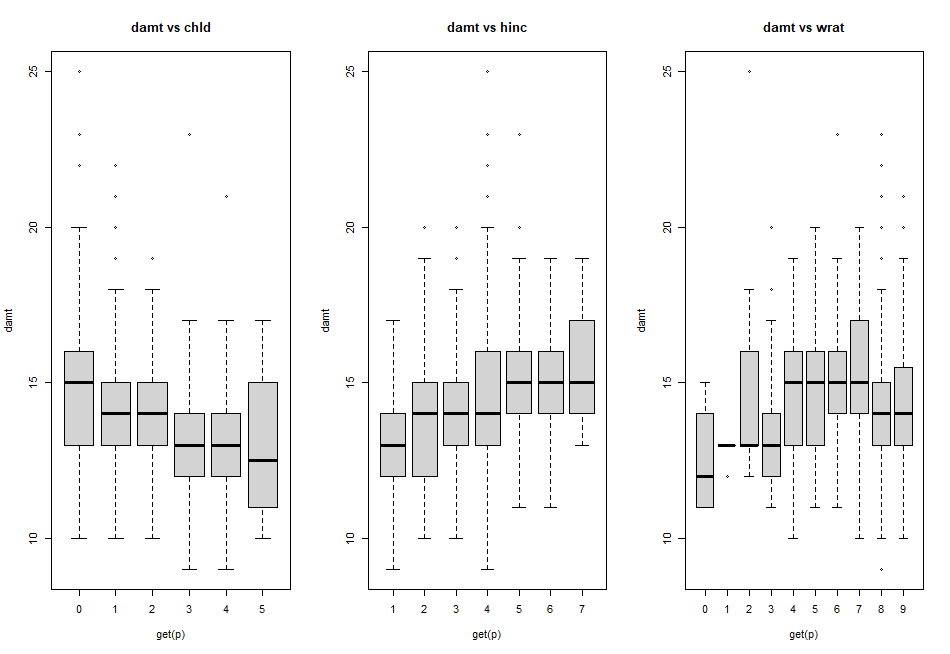
The summaries of these categories for the training data are described in the table below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Min* | *1st Qu* | *Median* | *Mean* | *3rd Qu* | *Max* | *SD* |
| *home* | *0.000* | *1* | *1* | *0.8833* | *1* | *1* | *0.3206* |
| *chld* | *0.000* | *0* | *2* | *1.577* | *3* | *5* | *1.4004* |
| *hinc* | *1.000* | *3* | *4* | *3.946* | *5* | *7* | *1.4007* |
| *genf* | *0.000* | *0* | *1* | *0.6049* | *1* | *1* | *0.4874* |
| *wrat* | *0.000* | *6* | *8* | *7.053* | *9* | *9* | *2.3257* |

The correlation matrix for these variables shows that home and chld have promising relationships with donr and damt. Home has a correlation of 0.29 with donr and 0.28 with damt, while chld has a -0.54 correlation with donr and a -0.57 correlation with damt. Wrat is also promising as it has a correlation of 0.17 with donr and 0.14 with damt.



The bar charts for chld, hinc, and wrat are below. These charts show that there is a clear downward trend in donations with more children and a clear upward trend with household income and wealth rating. These trends will be explored in the data preparation section.



#### Prior Gift Variables

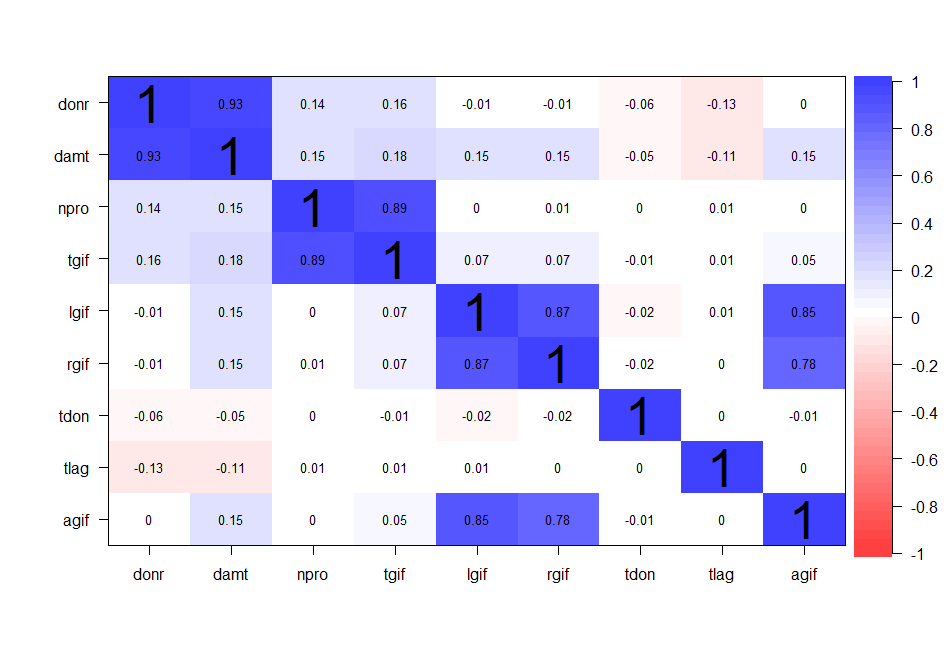
This section will focus on prior outreach attempts and prior donations. This will include the following variables:

|  |  |
| --- | --- |
| NPRO | *Lifetime number of promotions received to date* |
| TGIF | *The dollar amount of lifetime gifts to date* |
| LGIF | *The dollar amount of the largest gift to date* |
| RGIF | *The dollar amount of the most recent gift* |
| TDON | *Number of months since the last donation* |
| TLAG | *Number of months between first and second gift* |
| AGIF | *The average dollar amount of gifts to date* |

The summaries of these categories for the training data are described in the table below:

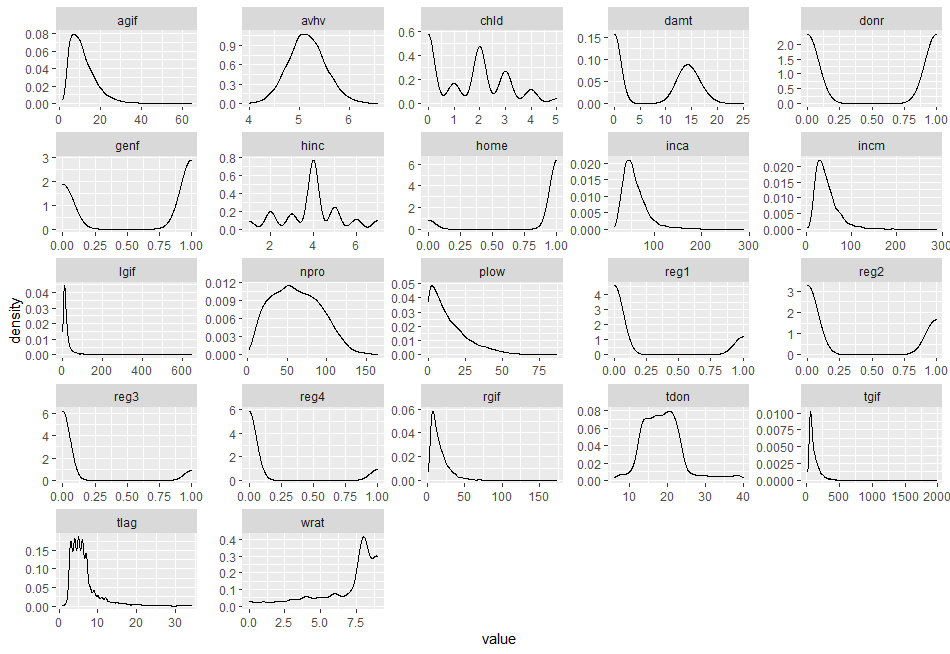
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Min* | *1st Qu* | *Median* | *Mean* | *3rd Qu* | *Max* | *SD* |
| *npro* | *2.000* | *37* | *60* | *61.63* | *84* | *164* | *30.3848* |
| *tgif* | *25.000* | *65* | *91* | *116.7* | *143* | *1974* | *85.5097* |
| *lgif* | *3.000* | *10* | *15* | *23.19* | *25* | *642* | *30.6030* |
| *rgif* | *1.000* | *7* | *12* | *15.55* | *20* | *173* | *11.8427* |
| *tdon* | *6.000* | *15* | *18* | *18.81* | *22* | *40* | *5.2346* |
| *tlag* | *1.000* | *4* | *5* | *6.302* | *7* | *34* | *3.9052* |
| *agif* | *1.890* | *7* | *10* | *11.66* | *14.79* | *64.22* | *6.4996* |

The correlation matrix for these variables is not very promising. Npro, tgif, and tlag are the only variables with somewhat of a correlation to donr (0.14, 0.16, and -0.13 respectively). Damt has stronger correlations, but they are all below 0.2. There are also very high correlations between agif and lgif and rgif (0.85 and 0.78 respectively).

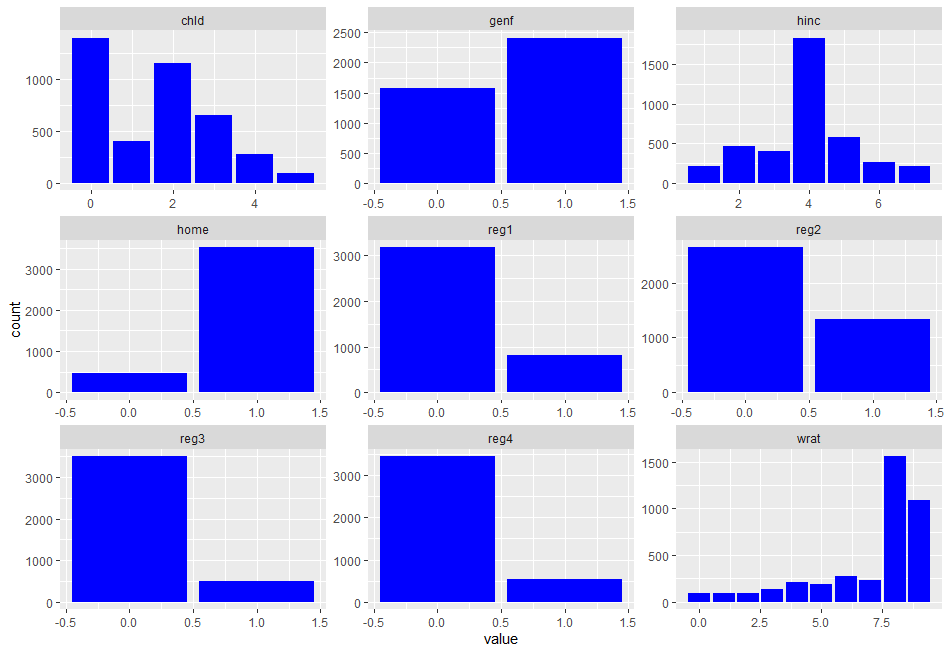
**

#### General Trends in Predictors

This section will cover the density plots and boxplots (for categorical variables) to show their distribution. The density plots show several notable right skews in the data that may need correction. Agif, inca, incm, lgif, rgif, tgif and tlag are notable examples of such a potential right skew. The only distributions that appear to be somewhat normal are avhv, damt (excluding zeroes), hinc (if it were smoothed), npro, and tdon.

**

The box plots for the categorical variables show similar information to the distribution plots but do a better job of capturing the bins.

**

### Data Preparation

The data for predictor and response variables in the training and validation sets was fully complete, not missing any observations. As such, we have retained all observations in an attempt to extract the most amount of information as is possible from the data. Also weighing into this decision to keep all observations was the lack of extreme outliers from the explanatory variables in the training set. In our opinion, the mean and median values for continuous quantitative values were close enough to each other, removing the need to eliminate observations from the data. For example, the median of the training observations for the average gift amount (agif) was 10.22 while the mean average gift amount was 11.66. Various other predictor variables such as average home value (avhv) showed similar behavior, with the median average home value of 171 and a mean average home value of 185.2. These figures suggest that many of the continuous economic data variables are skewed right, but the value of observations at the far right of the distribution does not skew the whole distribution enough to warrant removing outliers. Lastly, and more intuitively, the decision to retain observations at the right of the distribution was influenced by the fact that in the business of charity and fundraising, often the most important benefactors for a cause are those individuals with high income, high home value, and large past gifting amounts. Simply put, those with the most economic resources have the most to donate. As such, removing right outliers has the potential to systematically underestimate projected donation amounts in any subsequent modeling.

Given our decision to retain all observations, we do recognize the right skewness in many of the continuous economic variables. As a remedy, we have taken the log of the variables that showed more acute right skewness to produce a distribution that more closely resembles a normal distribution. Specifically, we have dropped the variables avhv, incm, inca, tgif, and agif from our analysis, replacing them with the log of each variable, respectively. We also create a new variable called financial wellness that is intended to proxy the financial health of a potential donor relative to that donor’s historic financial position. The variable is calculated as the most recent gift amount (rgif) divided by the largest gift amount (lgif) if lgif is not zero, zero otherwise. Larger ratios should indicate recent upward financial mobility and increased ability to donate whereas lower ratios should indicate recent financial hardship and less ability to donate.

Categorical and ranking variables also revealed some interesting results in our exploratory data analysis (EDA). Based on the training data, families with one child or less showed a vastly improved likelihood of donating. Knowing this, we have created a new dummy variable that returns true if a family has zero or one child, and false if the family has more than one child. Along the same school of thought, the seven categories of the hinc (household income) variable showed an increased likelihood of donating in the training data when the household falls into sections 3, 4, and 5 of income categories. We have created a new dummy variable that returns true if the hinc category is 3, 4, or 5 and false otherwise. Lastly, our EDA revealed the wrat (wealth rating) variable to have significant explanatory power in determining if an observation is a donor. Thus, we have created a third new dummy variable that returns true if the wealth rating is six or above and returns false for wealth categories less than six.

### Classification Models

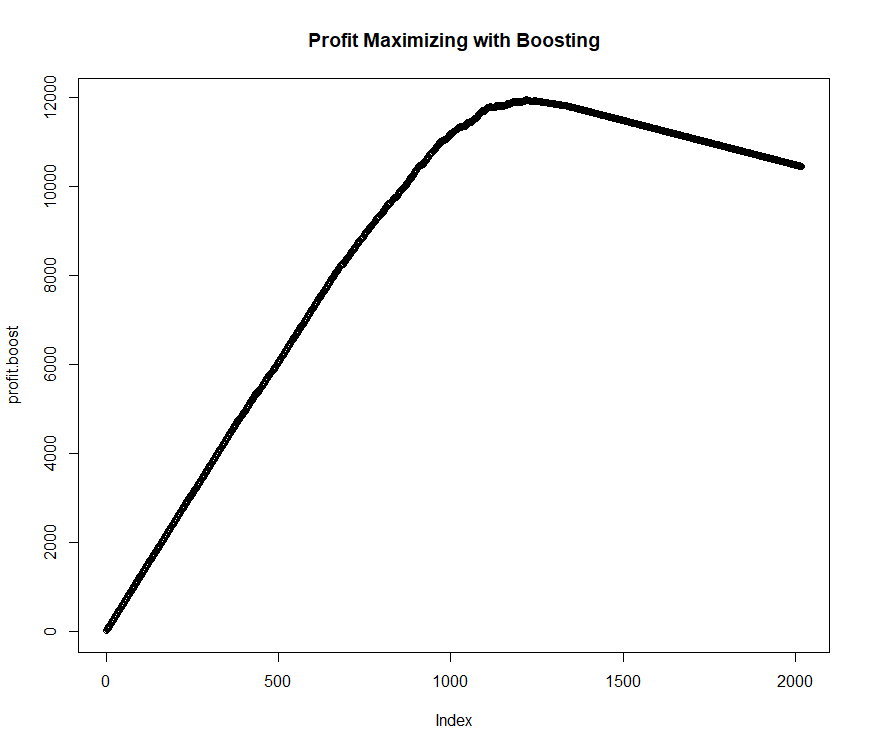
In this section, we developed various classification models on the variable DONR. There were 10 types of models considered for this section. All data were standardized before the models were run. All models were fit on the training data set and tested on the validation set to maximize the potential profit from a mailing using the model. Several models used hinc^2 as a factor as it increased the accuracy of the models. The rankings for the models are described in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Mail* | *Precision* | *Accuracy* | *Maximum Profit* |
| *BOOST* | *1218* | *0.8144* | *0.8845* | *$11948* |
| *KNN (using 100 neighbors)* | *1253* | *0.7909* | *0.8662* | *$11863.5* |
| *Logit GAM* | *1391* | *0.7168* | *0.8038* | *$11674.5* |
| *Support Vector Machine* | *1338* | *0.7377* | *0.8201* | *$11635.5* |
| *LDA* | *1367* | *0.7249* | *0.8097* | *$11635.5* |
| *LOG* | *1219* | *0.7957* | *0.8622* | *$11627* |
| *Support Vector Classifier* | *1378* | *0.7206* | *0.8038* | *$11599* |
| *Random Forest* | *1192* | *0.8079* | *0.8687* | *$11579.5* |
| *QDA* | *1293* | *0.7045* | *0.7820* | *$11261.5* |
| *Decision Tree* | *1027* | *0.8602* | *0.8702* | *$10720.5* |

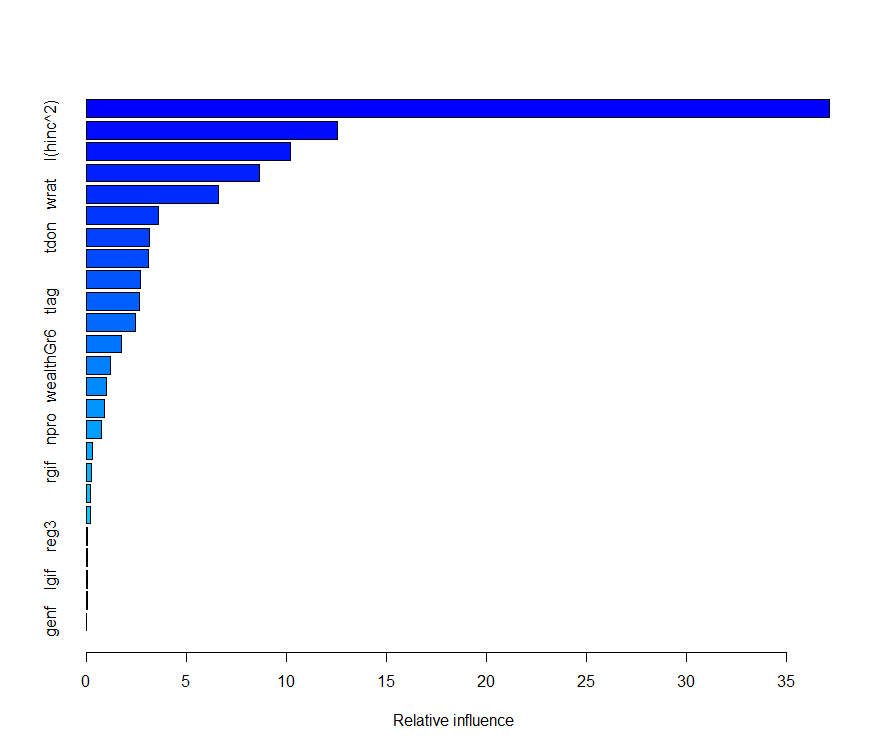
The Boosting model has the highest profit ($11948) and the best accuracy (88.45%). Interestingly, the Decision Tree model has the highest precision (86.02%), but the lowest maximum profit. This is likely due to the model also having the lowest amount of mail sent out. The boosting model maintains the second-highest precision out of all of the models. Overall the model is the best in the terms of the important metrics and is still strong on secondary metrics. The confusion matrix for the model and maximum profit chart for the model are below.

|  |  |  |
| --- | --- | --- |
| Validation  Boost Predictions | 0 | 1 |
| 0 | 793 | 7 |
| 1 | 226 | 992 |

The confusion matrix shows the high accuracy and precision of the model while also highlighting the nearly perfect precision rate of 99.30%



The maximum profit chart for the model shows that $12,000 is approximately the limit of the model based on changes in the amount of mail sent. As our model has $11,948 in profit, we are maximizing the potential profit.



The above chart shows that chld (37.1767 relative influence) has the most influence by a wide margin, followed by hinc^2 (12.5564 relative influence), reg2 (10.1990 relative influence), home (8.6808 relative influence), and wrat (6.6267 relative influence). After these five variables, there is a large drop in influence.

### 

### 

### 

### 

### 

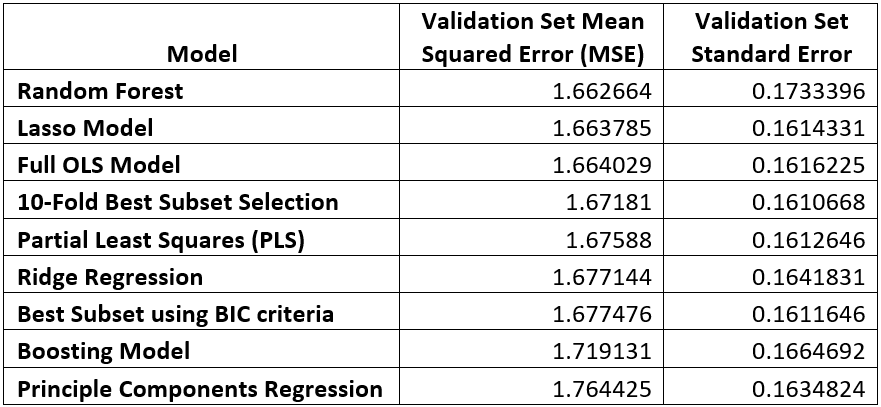
### 

### 

### 

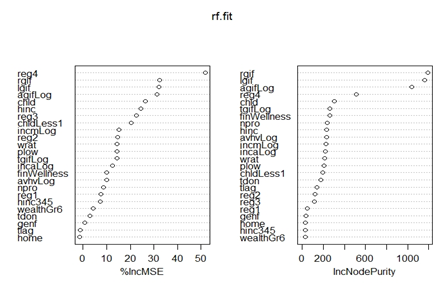
### Prediction Models

In this section, we developed various prediction models on the variable DAMT. The models considered were ordinary least squares (OLS) regression on the full set of variables, best subset selection with 10-fold cross-validation, principal components regression, partial least squares regression, ridge regression, the lasso model, boosting, best subset selection using BIC criteria, and a random forest model. All models were first fit on the training data and subsequently assessed and ranked based on the criteria of mean squared error resulting from predicting the validation data set.

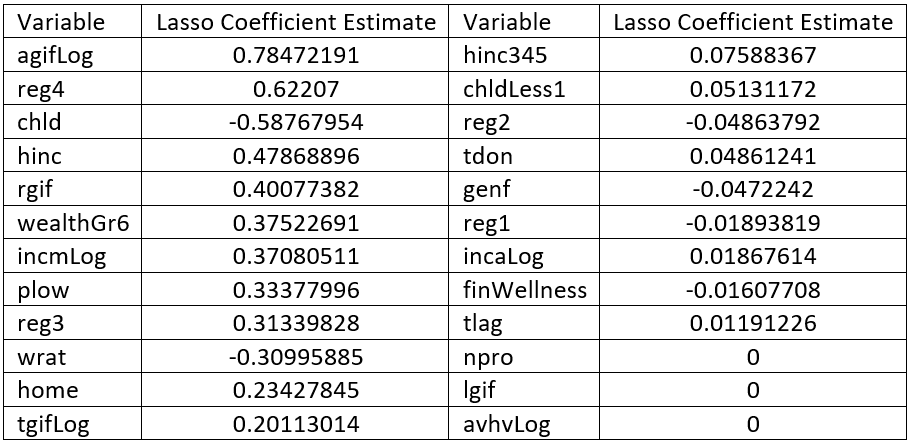


From the table above, you can see that the random forest model is the best fit on the validation data set, having the smallest MSE. However, the standard error from this model is significantly higher than all other models that were considered. These results indicate that the random forest model has low bias and high variance, possibly overfitting the data. The Lasso model is only marginally weaker in terms of mean square error ranking, yet, it has a substantially lower standard error in the context of these models that were considered. As such, the most robust model appears to be the Lasso model. However, given the criteria in the project to find the regression model with the lowest MSE, random forest is our best prediction model selection.

The chart below shows a scoring of variable importance in the random forest model, ranked by two metrics. The first metric is %IncMSE, which represents the percentage increase in MSE if that variable is removed from the model. The second is IncNodePurity, which represents the total decrease in node purity from that variable, averaged over all trees created in the random forest model. By both metrics, it is clear that the variables with the greatest explanatory power are reg4, rgif, lgif, chid, and agifLog.



The lasso model, with its strong MSE ranking, also exhibits a variable selection mechanism. Below are the coefficient estimates for the lasso model, ranked by absolute value. We can see the most important variables in the lasso model overlap with the random forest model, including: agifLog, reg4, child, hinc, and rgif.



### Results & Conclusion

We get the best results for boosting in the classification model and Random Forest in the prediction model using maximum profit and validation mean-squared error as the respective selection criteria. The boosting classification model predicts a maximum profit of $11948 based on sending out 1218 mailings. The prediction model has a validation set mean squared error of 1.662364 with a validation set standard error of 0.1733396.

The classification model predicts a response rate 0f 14.45%. This is 44.5% higher than the predicted response rate before the model was implemented (10%). The prediction model predicts an average donation of $14.22 across the test data set. This yields expected profits of

$14.22\*0.1445 - 2 = $0.05

If we restrict the mailing to only classification outcomes of expected donors, the average donation increases to $14.45. This yields the expected profits of

$14.45\*0.1445 - 2 = $0.09

While the above number may seem low, it is better than the $0.55 loss that was expected before the models were implemented. Overall the implementation of the boosting model for classifications and the random forest model for predictions have raised the maximum profit for sending out mailers significantly and the models even make it economically viable to send mailers out to a wider assortment of people.