Charity Course Project

Name: Young, Brent

MSDS 422 Section #: 55

Quarter: Spring 2018

Introduction

Problem

The purpose of the course project is to develop a machine learning model for a charitable organization using variables such as region, home ownership, number of children, household income, gender, neighborhood metrics, and donation metrics so that they can improve the cost-effectiveness of their direct marketing campaigns to previous donors. To achieve this, classification models using logistic regression, logistic regression GAM, LDA, QDA, k-nearest neighbors, neural networks, decision trees, bagging, random forest, boosting, and support vector machines will be fitted using the training data and then evaluated using the validation data (evaluation criteria: "maximum profit"). The final selected classification model will then be used to classify donor responses in the test dataset. In addition to this, we will also build a prediction model to predict expected gift amounts from donors. To accomplish this, prediction models using least squares regression, best subset selection with k-fold cross-validation, principal components regression, partial least squares, ridge regression, lasso, neural networks, decision trees, bagging, random forest, boosting, and support vector regression will be fitted using the training data and then evaluated using the validation data (evaluation criteria: "mean prediction error"). The final selected prediction model will then be used to predict donation amount responses in the test dataset.

Significance

The problem is significant/interesting because according to the organization's recent mailing records, the typical overall response rate is 10% and out of those who respond (donate) to the mailing, the average donation is $14.50. Additionally, each mailing costs $2.00 to produce and send (includes: a gift of personalized address labels and assortment of cards/envelopes). Therefore, it is not cost-effective to mail to everyone because the expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55. As a result, accurate predictions of likely donors using data from the most recent campaign will help ensure that the expected net profit is maximized.

Exploratory Data Analysis

Structure and Description of Training, Validation, and Test Datasets

The entire dataset consists of 24 variables and 8009 observations. The data is split into a 50/25/25: train/validation/test. For instance, there are 3984 training observations (50%), 2018 validation observations (25%), and 2007 test observations (25%). The training set will be used to fit the models, the validation set will be used to estimate prediction error for model selection, and the test set will be used for assessment of the prediction error of the final chosen model. Additionally, weighted sampling has been used so that the responders are over-represented on the training and validation samples and have approximately equal numbers of donors and non-donors. The response rate in the test sample has the more typical 10% response rate.

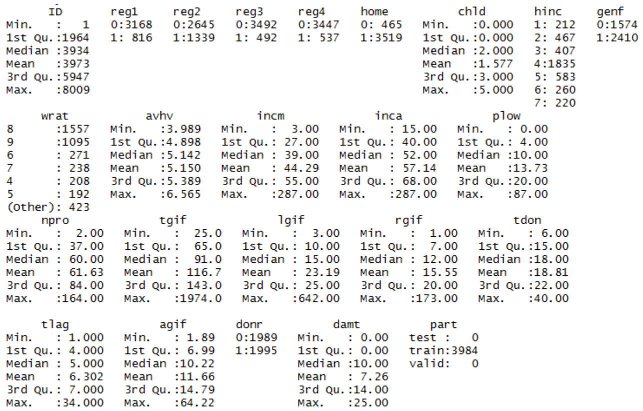
Variables

Variables such as ID (used for identification purposes) and PART (used to separate the datasets) will both be ignored. DONR represents our classification response variable (1 = Donor, 0 = Non-donor), while DAMT represents our prediction response variable (donation amount in $) and consists of the records for donors only. There is a total of 1995 donors in the training dataset and 999 donors in the validation dataset. It’s also important to note that both DONR and DAMT variables are set to “NA” for the test set (2007 NA’s). The other 20 variables consist of region, home ownership, number of children, household income, gender, neighborhood metrics, and donation metrics. All of these variables have been standardized to have a mean of 0 and standard deviation of 1 in the training, validation, and test datasets. Lastly, all variables with the exception of PART are integer variables. However, variables such as REG, HOME, HINC, GENF, WRAT, and DONR were changed to factor variables on the training dataset for EDA purposes and were then changed back to integers for model building purposes.

EDA for Classification Models

Descriptive Statistics

Observations: Figure 1 shows summary statistics of the variables in the training dataset so that we can check for missing values, outliers, distributions, etc. The data shows that there is an equal amount of donors and non-donors. Additionally, the mean donation amount is $7, median is $10, minimum is $0, and max is $25. There also appears to be some outliers for all the numeric variables, except CHLD. Furthermore, all the numeric variables have a right skew, except for AVHV, which has a normal distribution since it was log transformed as part of the full dataset. These distributions and outliers were



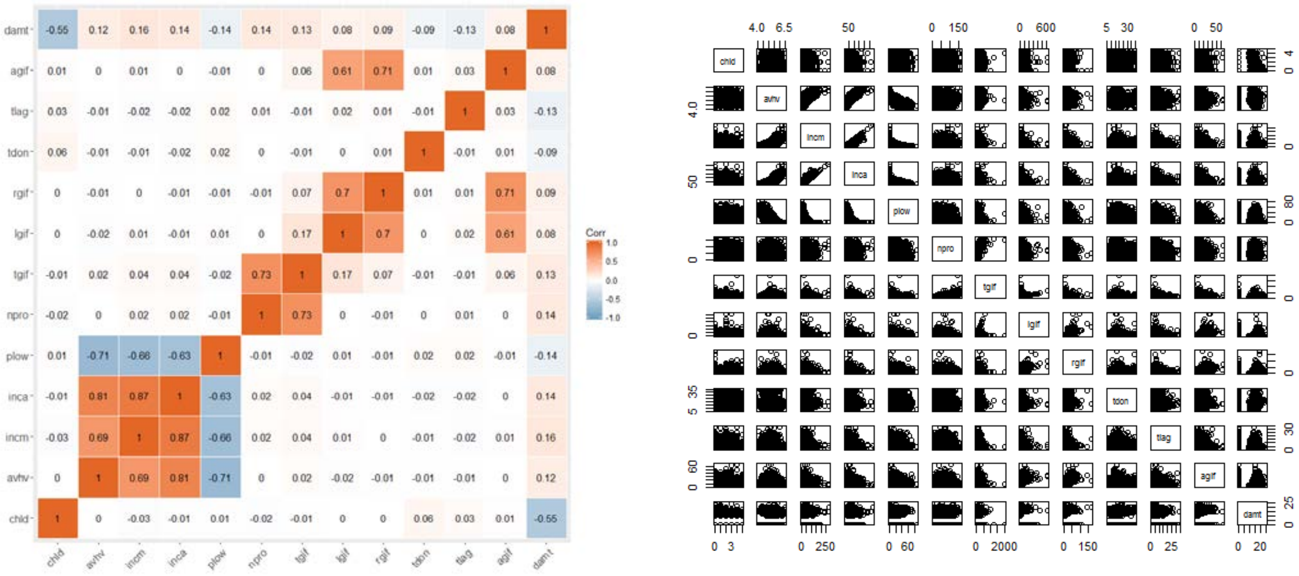
```
       ID          reg1      reg2      reg3      reg4      home          chld         hinc      genf
 Min.   :   1   0:3168   0:2645   0:3492   0:3447   0: 465   Min.   :0.000   1: 212   0:1574
 1st Qu.:1964   1: 816   1:1339   1: 492   1: 537   1:3519   1st Qu.:0.000   2: 467   1:2410
 Median :3934                                                Median :2.000   3: 407
 Mean   :3973                                                Mean   :1.577   4:1835
 3rd Qu.:5947                                                3rd Qu.:3.000   5: 583
 Max.   :8009                                                Max.   :5.000   6: 260
                                                                             7: 220
      wrat           avhv            incm              inca              plow
 8      :1557   Min.   :3.989   Min.   :  3.00   Min.   : 15.00   Min.   : 0.00
 9      :1095   1st Qu.:4.898   1st Qu.: 27.00   1st Qu.: 40.00   1st Qu.: 4.00
 6      : 271   Median :5.142   Median : 39.00   Median : 52.00   Median :10.00
 7      : 238   Mean   :5.150   Mean   : 44.29   Mean   : 57.14   Mean   :13.73
 4      : 208   3rd Qu.:5.389   3rd Qu.: 55.00   3rd Qu.: 68.00   3rd Qu.:20.00
 5      : 192   Max.   :6.565   Max.   :287.00   Max.   :287.00   Max.   :87.00
 (Other): 423
      npro              tgif              lgif              rgif              tdon
 Min.   :  2.00   Min.   :  25.0   Min.   :  3.00   Min.   :  1.00   Min.   : 6.00
 1st Qu.: 37.00   1st Qu.:  65.0   1st Qu.: 10.00   1st Qu.:  7.00   1st Qu.:15.00
 Median : 60.00   Median :  91.0   Median : 15.00   Median : 12.00   Median :18.00
 Mean   : 61.63   Mean   : 116.7   Mean   : 23.19   Mean   : 15.55   Mean   :18.81
 3rd Qu.: 84.00   3rd Qu.: 143.0   3rd Qu.: 25.00   3rd Qu.: 20.00   3rd Qu.:22.00
 Max.   :164.00   Max.   :1974.0   Max.   :642.00   Max.   :173.00   Max.   :40.00

      tlag              agif          donr          damt            part
 Min.   : 1.000   Min.   : 1.89   0:1989   Min.   : 0.00   test :   0
 1st Qu.: 4.000   1st Qu.: 6.99   1:1995   1st Qu.: 0.00   train:3984
 Median : 5.000   Median :10.22            Median :10.00   valid:   0
 Mean   : 6.302   Mean   :11.66            Mean   : 7.26
 3rd Qu.: 7.000   3rd Qu.:14.79            3rd Qu.:14.00
 Max.   :34.000   Max.   :64.22            Max.   :25.00
```

Figure 1

validated using histograms, boxplots, and quantiles as well. However, I decided not to handle/truncate the outliers because it did not improve my models much (or made them worse) when I tested them on the validation dataset. In regards to the qualitative variables, most of the donors/non-donors are from REG1, REG2, own a home, have a household income of category 4, are female, and have a wealth rating of 8/9. These insights were also similar to what I saw in the full dataset as well.

Numeric Variables

Figure 2: Correlation and Scatterplot Matrix



Observations: Figures 2 shows a correlation and scatterplot matrix of the variables that were included in the dataset (along with donation amount). This gives us an idea of the most promising predictor variables based on the predictors that are most correlated with DAMT and therefore DONR's. The following variables are most correlated with DAMT (near +/-0.14 or greater): CHLD, INCM, INCA, PLOW, NPRO, TGIF, and TLAG. Overall, the results show that CHLD is the most strongly negatively correlated to DAMT, which makes sense given that the more children someone has to support, the less money he/she has to donate (e.g., children are expensive). Furthermore, as average/median family income in a potential donor's neighborhood increases, percent categorized as low income in a potential donor's neighborhood decreases, lifetime promotions received to date increases, total dollar amount of lifetime gifts to date increases, and number of months since last donation decreases, donation amount tends to increase (and therefore will donate). Furthermore, most of the income related variables are strongly correlated with each other (e.g., INCM and INCA), which makes sense. As a result, this provides evidence to only include one and not both in some of our models. Additionally, the donation gift amount metric variables are strongly correlated with each other as well. Interestingly, NPRO and TGIF are strongly correlated with each other, which makes sense given that people that receive more promotions over the course of their lifetime are more likely to make more money and therefore donate more.

Figure 3a & 3b: Boxplot of CHLD vs. DONR & TDON vs. DONR

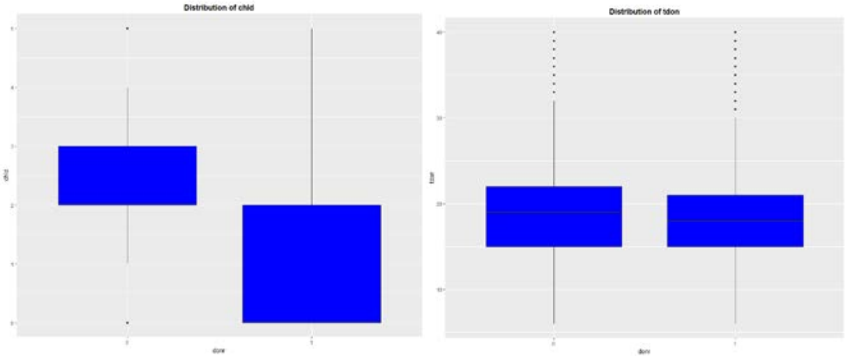


Figure 3a

Figure 3b

Observations: Figure 3a shows a boxplot of CHLD vs. DONR, so that we can compare the median differences and variability between the numeric variable and donors vs. non-donors. The results show that the less children someone has, the more likely they are to donate (vice versa). Figure 3a also provides evidence that CHLD is a strong predictor to include in our models since the median difference between donors and non-donors is wide. This was also confirmed in our correlation matrix. Figure 3b shows a boxplot of TDON vs. DONR, the results show that as the months since last donation begin to increase, the less likely they are to donate. The median difference between donors and non-donors are also close, but there are subtle differences. As a result, this provides evidence of possibly including this variable in our models.

Figure 4a & 4b: Boxplot of INCM vs. DONR & PLOW vs. DONR

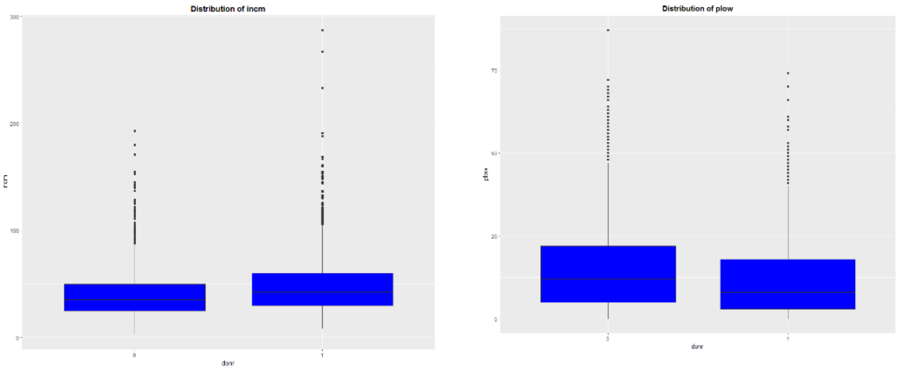


Figure 4a

Figure 4b

Observations: I also conducted additional boxplots of DONR (x-axis) vs. the other numeric variables (Y-axis) and found that INCM (figure 4a), INCA, PLOW (figure 4b), NPRO, and TLAG have median differences between donor and non-donors, while the rest of the numeric variables did not. Interestingly, these same variables listed above also had correlations of around +/-0.14 to DAMT in the correlation matrix (with the exception of TGIF). As a result, this provides strong evidence of possibly including INCM, PLOW, NPRO, and TLAG in our models.

Qualitative Variables

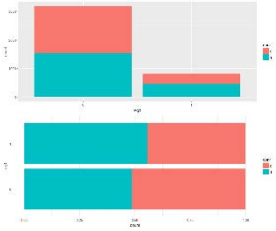


Figure 5: REG1

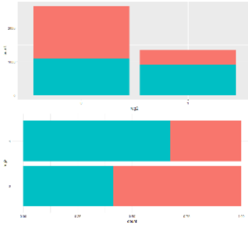


Figure 6: REG2

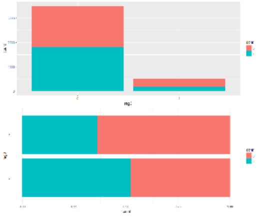


Figure 7: REG3

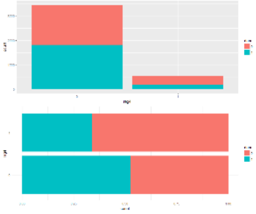


Figure 8: REG4

Observations: Figures 5 to 8 show bar plots of REG1, REG2, REG3, and REG4 by donor (blue) vs. non-donor (red). The data shows that people that live in the more populated regions such as REG1 and REG2 tend to donate, while those who live in the less populated regions such as REG3 and REG4 tend to donate less frequently. As a result, this provides evidence of possibly including REG1 and REG2 in our models.

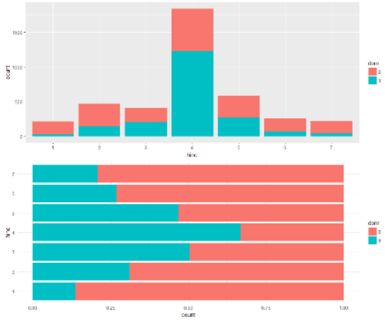


Figure 9: HINC

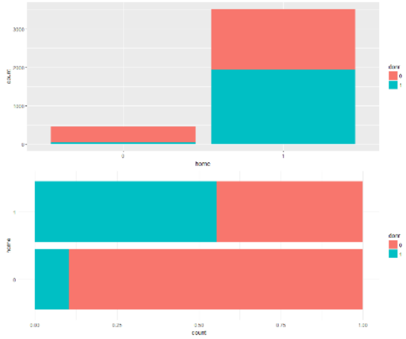


Figure 10: HOME

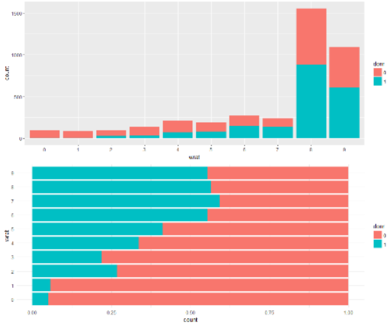


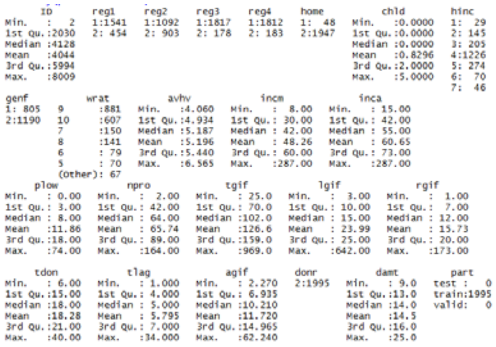
Figure 11: WRAT

Observations: Figures 9 to 11 shows bar plots of HINC, HOME, and WRAT by donor (blue) vs. non-donor (red). The data shows that people that have a household income of category 4, tend to donate more than the other categories, while those who have a household income of category 3 or 5 tend to donate around 50% of the time. Furthermore, the data shows that people who own a home tend to donate more frequently than those who do not own a home. Additionally, people that have a wealth rating of 6 or higher tend to donate more frequently than those who have a wealth rating of 5 or less. Lastly, I also looked at GENF as well using bar plots by donor vs. non-donor, but the data showed that there wasn’t much differentiation between female donors (49%) vs. non-donors (51%) and male donors (51%) vs. non-donors (49%). Therefore, this provides evidence that HINC, HOME and WRAT should possibly be included in our models, whereas GENF should not be included in some of our models.

EDA for Prediction Models

Descriptive Statistics

Observations: Figure 12 shows summary statistics of the variables in the training dataset (donors only) so that we can check for missing values, outliers, distributions, etc. The data shows that the mean donation amount for donors are $14.50, median is $14, minimum is $9, and max is $25. There also appears to be outliers for all the numeric variables, except CHLD. Furthermore, all the numeric variables have a right skew, except for AVHV, which has a normal distribution since it was log transformed as part of the full dataset. These distributions and outliers were validated using histograms and boxplots as well. In regards to the qualitative variables, most of the donors are from REG 1, REG 2, own a home, have a household income of category 4, are female, and have a wealth rating of 8/9. These insights were also similar to what we saw in the EDA for our classification models.

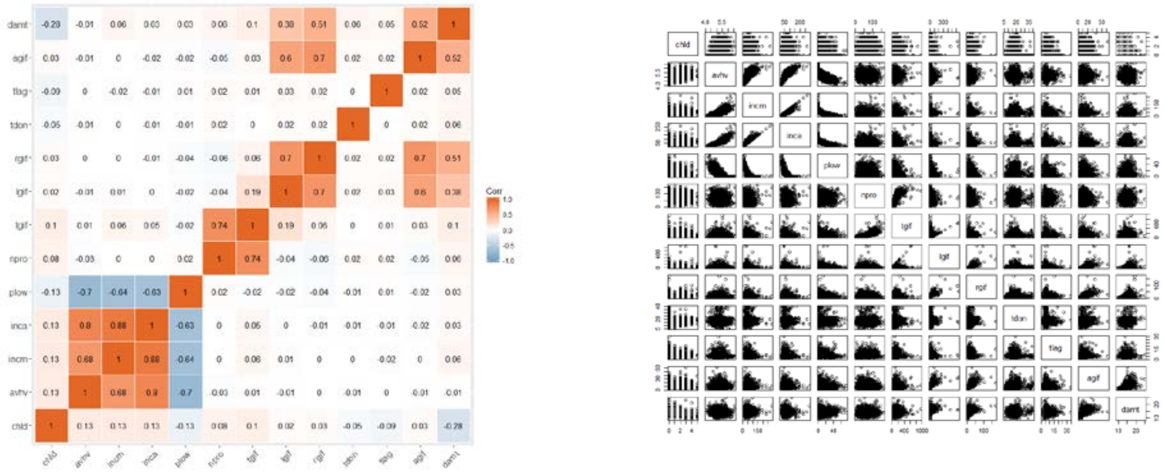


```
      ID          reg1      reg2      reg3      reg4      home         chld            hinc
 Min.   :   2   1:1541    1:1092    1:1817    1:1812    1:  48    Min.   :0.0000    1:  29
 1st Qu.:2030   2: 454    2: 903    2: 178    2: 183    2:1947    1st Qu.:0.0000    2: 145
 Median :4128                                                     Median :0.0000    3: 205
 Mean   :4044                                                     Mean   :0.8296    4:1226
 3rd Qu.:5994                                                     3rd Qu.:2.0000    5: 274
 Max.   :8009                                                     Max.   :5.0000    6:  70
                                                                                    7:  46
 genf          wrat           avhv             incm             inca
 1: 805    9      :881    Min.   :4.060    Min.   :  8.00    Min.   : 15.00
 2:1190    10     :607    1st Qu.:4.934    1st Qu.: 30.00    1st Qu.: 42.00
           7      :150    Median :5.187    Median : 42.00    Median : 55.00
           8      :141    Mean   :5.196    Mean   : 48.26    Mean   : 60.65
           6      : 79    3rd Qu.:5.440    3rd Qu.: 60.00    3rd Qu.: 73.00
           5      : 70    Max.   :6.565    Max.   :287.00    Max.   :287.00
           (Other): 67
      plow             npro              tgif             lgif              rgif
 Min.   : 0.00    Min.   :  2.00    Min.   : 25.0    Min.   :  3.00    Min.   :  1.00
 1st Qu.: 3.00    1st Qu.: 42.00    1st Qu.: 70.0    1st Qu.: 10.00    1st Qu.:  7.00
 Median : 8.00    Median : 64.00    Median :102.0    Median : 15.00    Median : 12.00
 Mean   :11.86    Mean   : 65.74    Mean   :126.6    Mean   : 23.99    Mean   : 15.73
 3rd Qu.:18.00    3rd Qu.: 89.00    3rd Qu.:159.0    3rd Qu.: 25.00    3rd Qu.: 20.00
 Max.   :74.00    Max.   :164.00    Max.   :969.0    Max.   :642.00    Max.   :173.00

      tdon             tlag              agif          donr          damt          part
 Min.   : 6.00    Min.   : 1.000    Min.   : 2.270    2:1995    Min.   : 9.0    test :   0
 1st Qu.:15.00    1st Qu.: 4.000    1st Qu.: 6.935              1st Qu.:13.0    train:1995
 Median :18.00    Median : 5.000    Median :10.210              Median :14.0    valid:   0
 Mean   :18.28    Mean   : 5.795    Mean   :11.720              Mean   :14.5
 3rd Qu.:21.00    3rd Qu.: 7.000    3rd Qu.:14.965              3rd Qu.:16.0
 Max.   :40.00    Max.   :34.000    Max.   :62.240              Max.   :25.0
```

Figure 12

Numeric Variables

Figure 13: Correlation and Scatterplot Matrix



Observations: Figures 13 shows a correlation and scatterplot matrix of the variables that were included in the dataset (along with donation amount). This gives us an idea of the most promising predictor variables based on the predictors that are most correlated with DAMT. The following variables are most correlated with DAMT (near +/-0.38 or greater): AGIF, RGIF, LGIF, and CHLD. Overall, the results show that AGIF is the most strongly positively correlated to DAMT, which makes sense given that as donation amount increases, average dollar amount of gifts to date also increases. Furthermore, as dollar amount of most recent gift increases, dollar amount of largest gift to date increases, and the less children people have, donation amount also increases, which makes sense. Additionally, most of the income related variables are strongly correlated with each other (e.g., INCM and INCA), donation gift amount metric variables are strongly correlated with each other, and NPRO and TGIF are strongly correlated with each other. This is similar to what we saw in the EDA for our classification models.

Qualitative Variables

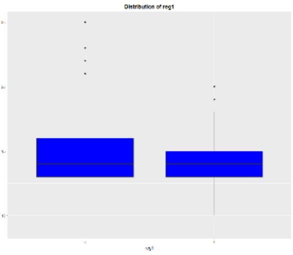


Figure 14 REG1

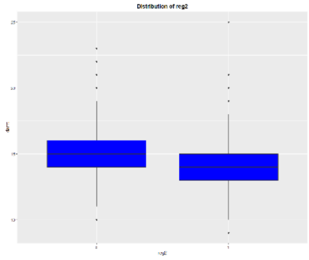


Figure 12 REG2

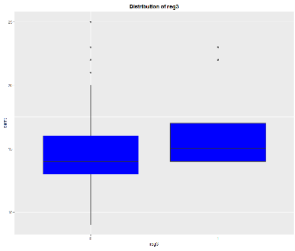


Figure 16 REG3

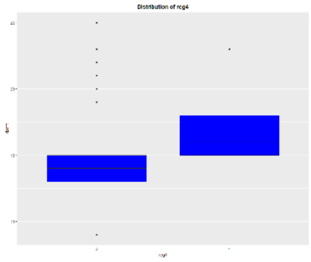


Figure 17 REG4

Observations: Figures 14 to 17 shows a boxplot of REG1 vs. DAMT, REG2 vs. DAMT, REG3 vs. DAMT, and REG4 vs. DAMT so that we can compare the median differences and variability between DAMT and the regions. The results show that REG2, REG3, and REG4 have the widest differences when compared to the median donation amount across all the other regions. For instance, people who live in REG1 tend to donate the median amount when compared to the other regions, while people who live in REG 2 tend to donate a lesser amount compared to the median across all the other regions. Furthermore, people who live in REG3 and REG4 tend to donate a higher amount compared to the median across all the other regions.
As a result, this provides evidence that REG2, REG3, and REG4 should possibly be included in our models.

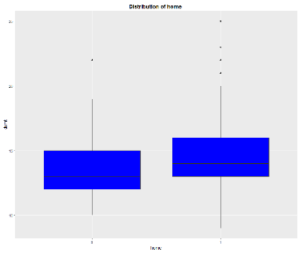


Figure 18 HOME

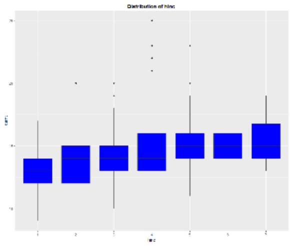


Figure 19 HINC

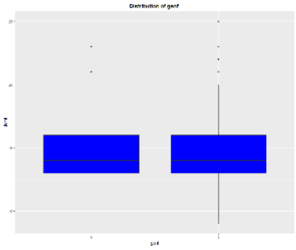


Figure 20 GENF

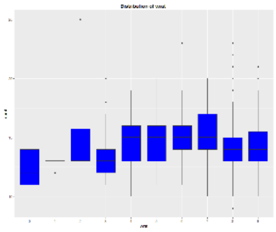


Figure 21 WRAT

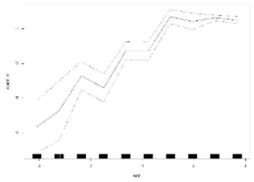
Observations: Figures 18 to 21 shows a boxplot of HOME vs. DAMT, HINC vs. DAMT, GENF vs. DAMT, and WRAT vs. DAMT so that we can compare the median differences and variability between DAMT and the rest of these qualitative variables. The results show that people who own a home tend to donate a higher amount compared to those who do not own a home. Additionally, the results show that there is a positive correlation between HINC and DAMT. As the household income categories increase, the donation amount also tends to increase as well. Furthermore, the results show that both males and females tend to give the same donation amounts. Lastly, the results show that there is no correlation between WRAT and DAMT. For instance, as wealth rating increases, the donation amount doesn't necessarily increase.

Formulation of Models

Classification Models

Logistic Regression: Logistic regression models the probability that the response variable belongs to a specific category (James, Witten, Hastie, & Tibshirani, 2013). As a result, using the glm function, we produced a logistic regression model of donr ~ reg1 + reg2 + home + chld + I(hinc^2) + wrat + incm + plow + npro + tdon + tlag. These variables were chosen using a combination of EDA, stepwise regression, and the varImp function from the caret package. The Analysis of Deviance table showed that all the variables that were included in this model were statistically significant, which illustrates that these variables improved the model. Additionally, the Analysis of Deviance table and varImp showed that CHLD, REG2, and HINC^2 impacted the model the most. The model produced the following performance metrics on the training dataset: AIC: 2191.22, BIC: 2266.702 and the following accuracy metrics and evaluation criteria on the validation dataset: accuracy: 0.8454, maximum profit: 11653.5 (donors: 1271.0), and AUC: 0.9426. This model was an improvement over the full model (baseline logistic regression model), which produced an AIC of 2201.584 and BIC of 2339.965 on the training dataset and an accuracy of 0.8375, maximum profit of 11642.5 (donors: 1291), and AUC of 0.9422 on the validation dataset.

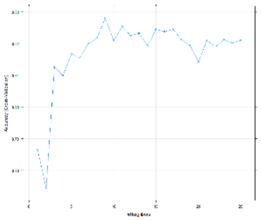
Logistic Regression GAM: Generalized additive models allow non-linear functions of each of the variables, while maintaining additivity and can be applied to classification problems (e.g., logistic regression) (James, et al., 2013). As a result, using the gam function, we then produced a logistic regression GAM model of donr~reg1 + reg2 + home + s(chld,4) + I(hinc^2) + s(wrat,9) + incm + plow + npro + s(tdon,4) + tlag. These variables were chosen using EDA and incorporating the variables from the best-performing logistic regression model. All of these variables were then given a smoothing spine with 4 degrees of freedom and compared against the best-performing logistic regression model using anova. The results showed that only CHLD, WRAT, and TDON were statistically significant. As a result, these variables were then given a smoothing spine with 4 degrees of freedom and were then tested with other variations of degrees of freedom using anova comparisons, the validation dataset, and using the plot function. After, conducting this analysis, it was determined that WRAT should have a degrees of freedom of 9, instead of 4. Evidence of using a degrees of freedom of 9 can be seen in the graph on the right, which shows that the confidence bands became narrower for WRAT. Additionally, the Analysis of Deviance table for the model showed that all the variables that were included in this model were statistically significant, which illustrates that these variables improved the model. Additionally, the Analysis of Deviance table showed that CHLD (df=4), REG2, and HINC^2 impacted the model the most, similar to the best performing logistic regression model. The model produced the following performance metrics on the training dataset: AIC: 1791.098, BIC: 1866.579 and the following accuracy metrics and evaluation criteria on the validation dataset: accuracy: 0.8707, maximum profit: 11829 (donors: 1234), and AUC: 0.9644. This provides evidence that this model is an improvement over the best performing logistic regression model.



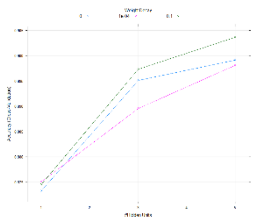
Linear Discriminant Analysis: LDA is very similar in form to logistic regression (distributions are assumed to be normal), except it models the distribution of the predictors separately in each of the response classes and then applies Bayes theorem (James, et al., 2013). Using the lda function, we then produced a linear discriminant analysis model of reg1 + reg2 + home + chld + I(hinc^2) + wrat + incm + plow + npro + tdon + tlag. These variables were chosen using EDA and incorporating the variables from the best-performing logistic regression model (see above). The model also produced the following accuracy metrics and evaluation criteria: accuracy: 0.8221, maximum profit: 11643.5 (donors: 1334.0), and AUC: 0.9418. This model was an improvement over the full model (baseline LDA model), which produced an accuracy of 0.8226, maximum profit of 11624.5 (donors: 1329), and AUC of 0.9414 on the validation dataset. However, LDA performed worse than logistic regression and logistic regression GAM.

Quadratic Discriminant Analysis: QDA is closely related to LDA, but it assumes that each class has its own covariance matrix (James, et al., 2013). Using the qda function, which incorporates similar syntax to LDA, we then produced a quadratic discriminant analysis model of reg1 + reg2 + home + chld + I(hinc^2) + wrat + incm + plow + npro + tdon + tlag. These variables were chosen using EDA and incorporating the variables from the best-performing logistic regression model. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.7592, maximum profit: 11274 (donors: 1439), and AUC: 0.9214 on the validation dataset. However, QDA performed worse than logistic regression, logistic regression GAM, and LDA.

K-Nearest Neighbors: KNN is a non-parametric method that applies Bayes rule by classifying a given observation to the class with the highest estimated probability (James, et al., 2013). Using the knn function, we then produced a KNN model with K=9. Prior to creating the KNN model, we created a training and validation matrix (removing DONR) specifically for the KNN model. Using the function trainControl in the caret package (grid search), K=9 was advised (see graph on the right). The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.8147 and maximum profit: 11025.5 (donors: 1237) on the validation dataset. However, KNN produced the lowest maximum profit when compared to logistic regression, logistic regression GAM, LDA, and QDA.



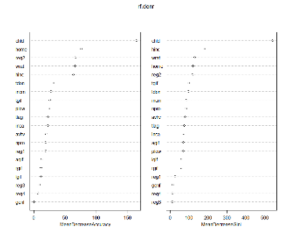
Neural Network: Using the avNNet function, we then produced a neural network model of reg1 + reg2 + home + chld + I(hinc^2) + wrat + incm + plow + npro + tdon + tlag. *Note: avNNet was used over nnet since it incorporates model averaging (default= 5 repeats), which is often advised for neural network due to the "randomness" of the model*. The variables above were chosen using EDA and incorporating the variables from the best-performing logistic regression model. Additionally, using the variables from the best-performing logistic regression model makes sense given that neural network is essentially a bunch of logistic regressions, fed into a multinomial logit model. I also incorporated 5 hidden layers into the model with a decay=0.10 and maxit=2000, which was determined using a grid search (see graph on the right). The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.870664, maximum profit: 11871 (donors: 1242), and AUC:



0.9638 on the validation dataset. As a result, according to these metrics and evaluation criteria, neural network outperformed logistic regression, logistic regression GAM, LDA, QDA, and KNN.

Decision Tree: A decision tree is a tree-based method that involves stratifying or segmenting the predictor space into a number of simple regions (James, et al., 2013). Decision trees are simple and useful for interpretation, but are generally not as competitive when it comes to prediction accuracy as other supervised learning approaches and suffer from high variance (James, et al., 2013). Prior to creating the decision tree model, we created a separate training and validation matrix for all the tree-based methods, since it requires DONR to be coded as a factor variable. We then used the tree function to produce a decision tree model with 15 predictor variables, after cross-validation helped eliminate 5 predictor variables. However, the accuracy and profit remained the same. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.8484, AUC: 0.909, and maximum profit: 11149 (donors: 1168) on the validation dataset. However, although the decision tree accuracy score was decent, it produced the second lowest maximum profit when compared to the other models.

Bagging: Bagging is a general-purpose procedure for reducing the variance and increasing the accuracy of a statistical method by taking repeated samples from the training dataset, with the cost of reducing interpretability (James, et al., 2013). As a result, using the randomForest function (mtry=20, ntree=100), a bagged decision tree model was produced using all 20 predictor variables. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.884, AUC: 0.950, and maximum profit: 10959 (donors: 1031) on the validation dataset. However, although the decision tree accuracy score improved, it produced the lowest maximum profit when compared to the other models.



Random Forest: Random forest provides an improvement over bagged trees by incorporating a small tweak that decorrelates the trees (e.g., forces each split to only consider a subset of predictors and will not consider strong predictors so that other predictors will have more of a chance (James, et al., 2013)). As a result, using the randomForest function (mtry=4, ntree=400), a random forest model was produced using all 20 predictor variables. Using the variable importance function, the plot on the right shows that CHLD, HINC, WRAT, HOME, and REG2 are the most important variables. This is similar to what we saw in our EDA. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.893, AUC: 0.960, and maximum profit: 11266 (donors: 1066.0) on the validation dataset. This was an improvement over a basic decision tree and bagging. The accuracy score was also the highest compared to all the models we've fitted thus far. However, maximum profit was lower than logistic regression, logistic regression GAM, LDA, QDA, and neural network.

Boosting: Boosting provides another approach for improving the predictions resulting from a decision tree by fitting each tree on an altered version of the original dataset (James, et al., 2013). In other words, trees are grown sequentially (e.g., each tree is grown using information from previously grown trees). As a result, using the gbm function, a boosted decision tree model was produced using all 20 predictor variables. I also incorporated n.trees =200, shrinkage=0.1, and depth=2, which was determined using a grid search. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.8786, AUC: 0.972, and maximum profit: 11955.5 (donors: 1236.0) on the validation dataset. As a result, according to these metrics and evaluation criteria, boosting outperformed all the other models.

Support Vector Machines: Support vector machine (SVM) is often considered one of the best "out of the box" classifiers (James, et al., 2013). As a result, using the caret package (method = svmLinear), a support vector classifier model, which is a *linear* classifier that allows some training observations to be misclassified in order to improve classification for the remaining observations was produced using all 20 predictor variables (James, et al., 2013). I also incorporated a cost of 0.01, which was determined using a grid search. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.837, maximum profit: 10522.5 (donors: 1822.0) on the validation dataset. I then decided to fit a support vector machine with a radial kernel, which is a *non-linear* classifier that cannot be perfectly separated with a hyperplace and is an extension of the support vector classifier, but uses kernels to enlarge the feature space (James, et al., 2013). I also incorporated a cost of 10 and gamma of 0.025, which was determined using a grid search as well. The model produced the following accuracy metrics and evaluation criteria: accuracy: 0.8791, maximum profit: 11023 (donors: 1057) on the validation dataset. As a result, the support vector machine produced much better results than the support vector classifier model. However, according to maximum profit, support vector machine performed worse than all the models, except bagging.

Results: Performance/Accuracy of Classification Models

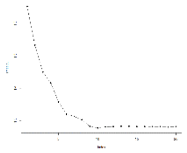
Model Name	Donors	Max Profit	Accuracy	AUC
Logistic Regression (Starter 1)	1291	11642.5	0.8375	0.942
Logistic Regression 2	1271	11653.5	0.8454	0.943
Logistic Regression GAM	1234	11829	0.8707	0.964
Linear Discriminant Analysis (Starter 1)	1329	11624.5	0.8226	0.941
Linear Discriminant Analysis 2	1334	11643.5	0.8221	0.942
Quadratic Discriminant Analysis	1439	11274.0	0.7592	0.921
K-nearest Neighbors	1237	11025.5	0.8147	N/A
Neural Network	1242	11871.0	0.8707	0.964
Decision Tree	1168	11149.0	0.8484	0.909
Bagging	1031	10959.0	0.884	0.950
Random Forest	1066	11266	0.893	0.960
Boosting	1236	11955.5	0.8786	0.972
Support Vector Machines	1057	11023	0.8791	N/A

Figure 22

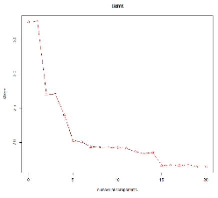
Observations: Figure 22 shows the performance/accuracy metrics and evaluation criteria for the following models in the validation dataset: logistic regression, logistic regression GAM, LDA, QDA, k-nearest neighbors, neural networks, decision trees, bagging, random forest, boosting, and support vector machines. The results show that the boosting model produced the highest max profit, while the bagging model produced the lowest max profit out of all the models. The boosting model also produced high accuracy and had the highest AUC as well. Given that the boosting model is the clear winner, I then applied the model to the test dataset, while also incorporating the oversampling adjustment. This adjustment is needed since the validation response rate is 0.5, but the test data response rate is 0.1, the optimal mailing rate in the validation data needs to be adjusted before we applied it to the test data. The predictions for the DONR variable after the oversampling adjustment was applied showed 1707 non-donors and 300 likely donors. Therefore, based on this model, we'll mail to the 300 highest posterior probabilities.

Prediction Models

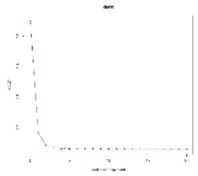
Multiple Linear Regression: Multiple linear regression uses predictors to predict the dependent variable (Y) (James, et al., 2013). As a result, using the lm function, we produced a multiple linear regression model of damt ~ reg2 + reg3 + reg4 + home+ chld + hinc + incm + plow + npro + rgif + agif. These variables were chosen using a combination of EDA, stepwise regression, and the varImp function from the caret package. The residual standard error of 1.275, showed us that when predicting DAMT, one standard error = 1.275. The adjusted r-squared value of 0.5666, indicates that 56.66% of the variation in DAMT is explained by the predictor variables. The adjusted r-squared value is slightly lower than the full model (0.5679), but the model is more parsimonious. The model also produced the following evaluation criteria on the training dataset: AIC: 6643.565, BIC: 6716.344 and mean prediction error: 1.846847 and standard error: 0.1682736 on the validation dataset. This model was an improvement over the full/baseline model, which produced an AIC of 6646.579 and BIC of 6769.743 on the training dataset and a mean prediction error of 1.867523 and standard error of 0.1696615 on the validation dataset.



Best Subset Selection with k-fold Cross-validation Model: Best subset selection involves identifying a subset of the predictors that we believe to be related to the response variable, which is then fit using least squares regression on the reduced set of variables (James, et al., 2013). As a result, using the regsubsets function, variables were then chosen using 10-folds cross-validation. The figure to the right shows us that using a 10 variable model is best, since the RMSE is lowest at 10. As a result, in comparison to the MLR model above, regsubsets chose the same variables that were included in the MLR model with the exception that it did not include REG2. The residual standard error of 1.275, shows us that when predicting DAMT, one standard error = 1.275. The adjusted r-squared value of 0.5663, indicates that 56.63% of the variation in DAMT is explained by the predictor variables. The adjusted r-squared value is slightly lower than the MLR model (0.5666), but is slightly more parsimonious. The model also produced the following evaluation criteria on the training dataset: AIC: 6643.918, BIC: 6711.099 and mean prediction error: 1.857947 and standard error: 0.1693538 on the validation dataset. As a result, this model performed slightly worse than the MLR model.



Principal Components Regression: Principal components regression is a dimension reduction method (unsupervised) that involves identifying linear combinations or directions that best represent the predictors (components) and then using them as predictors in a linear regression model that is fit using least squares (James, et al., 2013). As a result, using the pcr function, we then produced a principal components regression model of 15 principal components, which was chosen using cross-validation and the plot on the right. The model produced the following evaluation criteria: mean prediction error: 1.865497 and standard error: 0.1698902 on the validation dataset. As a result, according to the evaluation criteria, this model performed worse than multiple linear regression and best subset selection.



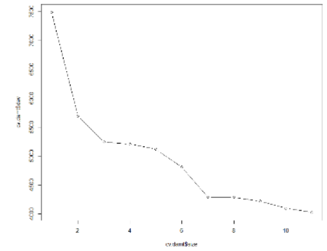
Partial Least Squares: Partial least squares is a dimension reduction method (supervised) that involves identifying a new set of features that are linear combinations of the original features and then fits a linear model using least squares using these new features that are also related to the response variable (James, et al., 2013). As a result, using the pls function, we then produced a partial least squares model with 3 components, which were chosen using cross-validation and the plot on the right. The model produced the following evaluation criteria: mean prediction error: 1.879345 and standard error: 0.1718698 on the validation dataset. As a result, this model performed worse than the MLR, best subset selection model, and PCR model.

Ridge Regression: Ridge regression is a regularization/shrinkage method that contains all the predictors and shrinks the coefficient estimates toward zero relative to the least squares estimates (James, et al., 2013). As a result, using the glmnet function (alpha=0), we then fitted a ridge regression model with lamda= 0.1107589 (tuning parameter). Additionally, prior to fitting the model, we also created a matrix for the training and validation datasets and then obtained lamda using 10 fold cross-validation. The model produced the following evaluation criteria: mean prediction error: 1.873279 and standard error: 0.1711228 on the validation dataset. As a result, this model performed worse than multiple linear regression, best subset selection, PCR, but better than partial least squares.

Lasso: The lasso, which is also a regularization/shrinkage method, is similar to ridge regression, but is a simpler model that performs variable selection, is easier to interpret, and shrinks some coefficients entirely to zero (James, et al., 2013). Using the same function and process as ridge regression, except with alpha=1, we then fitted a lasso model with lamda= 0.00877745 using 10 fold cross-validation. The model produced the following evaluation criteria: mean prediction error: 1.861709 and standard error: 0.1694331 on the validation dataset. As a result, this model performed better than PCR model, partial least squares, and ridge regression, but worse than multiple linear regression and best subset selection.

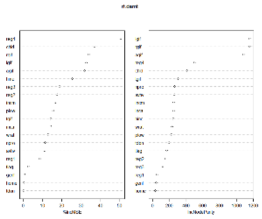
Neural Network: Using the avNNet function, we then produced a neural network model with all the predictor variables. I elected to use all the variables since neural network performed better compared to using a subset of variables. I also incorporated 5 hidden layers into the model with a decay=0.01 and maxit=500. The model produced the following evaluation criteria: mean prediction error: 1.466398 and standard error: 0.1620263 on the validation dataset. As a result, according to these metrics and evaluation criteria, neural network outperformed MLR, best subset selection model, PCR, partial least squares, ridge regression, and the lasso.

Decision Tree: Using the tree function, we then produced a decision tree model with 11 predictor variables after cross-validation helped eliminate 9 predictor variables. The model produced the following evaluation criteria: mean prediction error: 2.241075 and standard error: 0.1920681on the validation dataset. As a result, according to these metrics and evaluation criteria, the decision tree model performed the worst compared to all the other prediction models.



Bagging: To help improve performance, a bagged decision tree was then created using the randomForest function (m=20, ntree=100) with all 20 predictor variables. The model produced the following evaluation criteria, which was a huge improvement: mean prediction error: 1.699838 and standard error: 0.1759852 on the validation dataset. As a result, according to these metrics and evaluation criteria, the bagged decision tree model outperformed all the models, except for neural network.

Random Forest: To further enhance performance, a random forest model was then created using the randomForest function (mtry=4, ntree=400) with all 20 predictor variables. Using the variable importance function, the plot on the right shows that LGIF, RGIF, AGIF, CHLD, and REG4 are the most important variables. This was an improvement over a basic decision tree and bagging. The model produced the following evaluation criteria: mean prediction error: 1.663914 and standard error: 0.1733175 on the validation dataset. As a result, according to these metrics and evaluation criteria, random forest outperformed all the models, except for neural network.



Boosting: To further improve performance, a boosted decision tree model was produced using the gbm function, using all 20 predictor variables. I also incorporated n.trees =400, shrinkage=0.1, and depth=1, which was determined using a grid search. The model produced the following evaluation criteria: mean prediction error: 1.334194 and standard error: 0.1515649 on the validation dataset. As a result, according to these metrics and evaluation criteria, boosting outperformed all the other models.

Support Vector Regression: Support vector regression applies similar principles as support vector machine, but is considered the adapted form of SVM since the dependent variable is numerical, rather than categorical (James, et al., 2013). As a result, using the caret package (method = svmLinear), a support vector regression model (linear kernel) was produced using all 20 predictor variables. I also incorporated a cost of 0.01, which was determined using a grid search. The model produced the following evaluation criteria: mean prediction error: 1.865914 and standard error: 0.1785505 on the validation dataset. I then decided to fit a support vector regression with a radial kernel (method=svmRadial; non-linear kernel). I also incorporated a cost of 5 and gamma of 0.005, which was determined using a grid search as well. The model produced the following evaluation criteria: mean prediction error: 1.568629 and standard error: 0.1734832 on the validation dataset. As a result, the support vector regression with a radial kernel produced much better results than the support vector regression model with a linear kernel. In comparison to the other models, support vector regression outperformed all the models (evaluation criteria: “mean prediction error”), except neural network and boosting.

Results: Performance/Accuracy of Prediction Models

Model Name	Mean Prediction Error (MPE)	Standard Error
Multiple Linear Regression (Starter 1)	1.867523	0.1696615
Multiple Linear Regression (Starter 2)	1.867433	0.1696498
Multiple Linear Regression 3	1.846847	0.1682736
Best Subset Selection (K-folds CV)	1.857947	0.1693538
Principal Components Regression	1.865497	0.1698902
Partial Least Squares	1.879345	0.1718698
Ridge Regression	1.873279	0.1711228
The Lasso	1.861709	0.1694331
Neural Network	1.466398	0.1620263
Decision Tree	2.241075	0.1920681
Bagging	1.699838	0.1759852
Random Forest	1.663914	0.1733175
Boosting	1.334194	0.1515649
Support Vector Regression	1.568629	0.1734832

Figure 23

Observations: Figure 23 shows the performance/accuracy metrics and evaluation criteria for the following models in the validation dataset: least squares regression, best subset selection with k-fold cross-validation, principal components regression, partial least squares, ridge regression, lasso, neural networks, decision trees, bagging, random forest, boosting, and support vector regression. The results show that the boosting model produced the lowest mean prediction error (MPE), while the decision tree model produced the highest mean prediction error (MPE), out of all the models. Given that the boosting model is the clear winner, I then applied the model to the test dataset.

Summary of Test Results

```
chat.test
  0    1
1707  300
```

Mail to the 300 highest posterior probabilities.
Optimal test mailing rate is 0.1494

Summary Statistics ***Predicted donation amount for those who are expected to donate to the mailing***	
MEAN	14.34
MEDIAN	14.27
MAX	19.30
MIN	10.60

Observations: Overall, the boosting classification and prediction models predicts an expected profit of $4,303.29, with a mailing cost of $600 (300*$2). The optimal test mailing rate is 14.94%, which is higher than the typical response rate of 10%. Out of those who are expected to respond (donate) to the mailing, the average predicted donation is $14.34, which is close to the average donation of $14.50.

Conclusion

Recap

In section 1, we conducted an exploratory data analysis for classification and prediction models using correlation matrices, scatterplots, boxplots, histograms, summary statistics, etc. to help understand important characteristics and properties of the data that may be disguised by numerical summaries. The EDA revealed that reg1, reg2, home, chld, hinc, wrat, incm, plow, npro, tdon, and tlag were the most the promising predictors for classification models. Additionally, the EDA revealed that reg2, reg3, reg4, home, chld, hinc, incm, rgif, lgif, and agif were the most promising predictors for prediction models.

In section 2 and 3, we built 11 classification and 12 prediction models using different modeling methods (e.g., logistic regression, logistic regression GAM, LDA, QDA, KNN, neural network, tree-based methods, boosting, support vector machines, support vector regression, ordinary least squares, best subset selection, principal components regression, PLS, ridge regression, and the lasso). After evaluating the models using the performance and accuracy metrics, the results showed that boosting produced the best results for both classification and prediction.

Lastly, in section 4, the boosting classification and prediction model were both applied to the test dataset. The results showed that we should mail to the 300 highest posterior probabilities and should expect a profit of $4,303.29 and an average predicted donation for those who are expected to respond (donate) of $14.34.

Future Work

In regards to future work, there are three areas that could help improve my models. First, it would be beneficial to obtain additional demographic variables (e.g., education, job status, job type, etc.). Second, it could be helpful exploring different mixing of models using an ensemble approach since model diversity can help increase accuracy and performance. Third, it could be helpful to explore different transformations of the variables, interactions, and outlier management.

Learnings

In the end, I learned five primary things from building these models. First, I learned how to build different classification and prediction models using various methods. Second, I learned that it's really important to conduct a thorough EDA and that a lot can be learned from it. Third, I learned that trying different modeling approaches can result in better performance and to never settle on a model due to gut instinct. Fourth, I learned the importance of the confusion matrix, especially when we are trying to maximize donors or in this case profit (e.g., accuracy isn't the end all be all). Lastly, I learned how to use the caret package (e.g., tuning grids for models such as neural network, boosting, and KNN).

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. New York: Springer Science + Business Media, 2013.