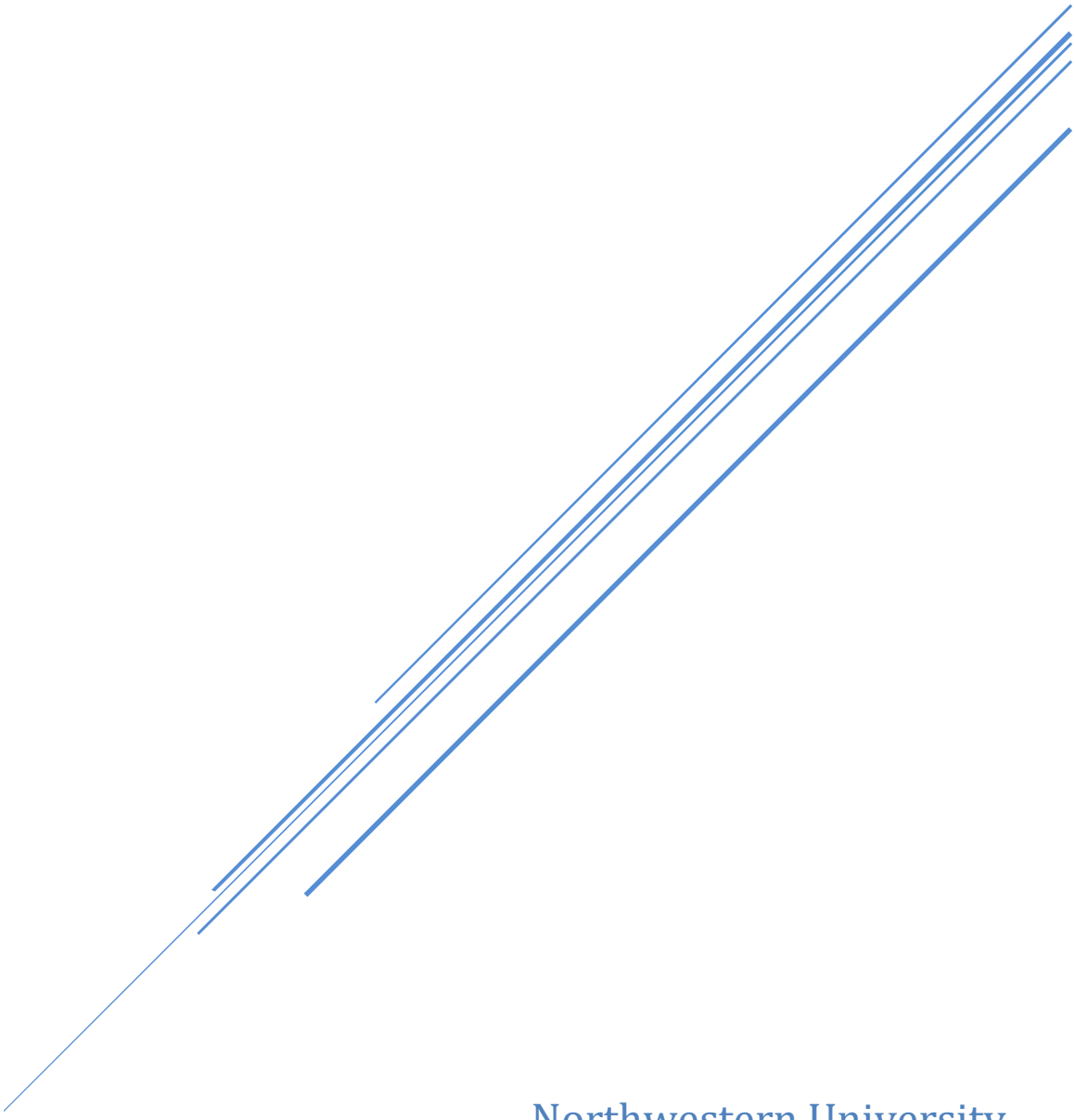
CHARITABLE ORGANIZATION CLASSIFICATION AND PREDICTION MODELS

YING CHENG, WILL CAVANAUGH, AND STEVEN SCHOIBER



Northwestern University
MSDS 422-DL

Introduction

The purpose of this project is to help a charitable organization maximize the potential successful outcomes of donations pertaining to their mail marketing efforts. According to their recent mailing records, the typical overall response rate is 10%. Of the 10% who donate, the average donation is $14.50. However, each mailing costs $2.00 to produce and send. Thus, it is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55. To help the charitable organization, succeed in their efforts, we will develop a classification model using data from the most recent campaign that can effectively captures likely donors so that the expected net profit is maximized. Additionally, we would also like to build a prediction model to predict expected gift amounts from donors.

Exploratory Data Analysis

Dataset Overview

The charity dataset consists of 8,009 observations that detail characteristics of sampled individuals. Across these 8,009 observations, we will divide the data into three sets: training data (3,984 observations), validation data (2,018), and test data (2,007). The variables comprising this dataset will allow us to predict whether or not a given individual will donate to our charity, and, if they do donate, also predict how much they will donate. Overall, there are 24 variables in this dataset, 17 of which will be considered for inclusion in our final models. The predictor variables are comprised of a mix between categorical and numeric variables.

Across the Exploratory Data Analysis, we will focus our attention on the training dataset, as this is the data that our predictive models will be built from. Some of the information outlined in the EDA, such as variable type and definitions, can also be applied to the same variables in the validation and test datasets.

Response Variable Exploration

In this analysis, there are two response variables that we are interested in studying - DONR and DAMT. In the following section, we will look at these two variables in an effort to better understand their make-up in hopes of having a better idea of what variables can be used to predict their values.

DONR

DONR is a classification response variable that indicates whether or not an observed individual donated to our charity. For this variable, a value of 1 indicates that the observed individual is a donor, while a value of 0 indicates that the observed individual is a non-donor. As we dive into the EDA for the predictor variables, we will explore each variable's relationship to DONR in order to gauge which variables may be good predictors for this response variable. The table below depicts the number of donors vs. non-donors within the dataset. In both the training and

validation datasets, the split between donors and non-donors is roughly even. This variable takes on a NA value in the test data, as this is the response variable we will be predicting.

DONR Values by Data Cut					
Training		Validation		Test	
0	1	0	1	0	1
1,989	1,995	1,019	999	NA	NA

DAMT

DAMT is a numeric response variable that indicates how much an individual donated to our charity. According to the summary statistics below, the giving amounts range from $0 (i.e. no giving) to $25. As indicated when observing the DONR variable above, there are 1,989 observations who did got donate to the charity which is represented by the 1,989 Zeros in the table below. The average gift, including Zeros, is $7.26 and the median value is $10. Given that this is a response variable we desire to predict, we will not perform any transformations or truncations for this variable. One initial metric that can be used to begin studying how each of the variables relate with our variable is correlation. According to the correlation matrix depicted below, none of the predictor variables share a very strong relationship with the response variable. At first glance, this indicates that a slightly more complex model featuring numerous variables will need to be used in order to accurately predict this variable.

	min	**1%**	**5%**	**25%**	**Mean**	**Median**
damt	0	0	0	0	7.26	10
75%	**95%**	**99%**	**Max**	**sd**	**Zeros**	**missing**
14	17	19	25	7.4	1989	0

Categorical Predictor Variable Exploration

Among the 17 predictor variables included within the dataset, five (5) of the included variables are categorical in nature. These variables will be summarized and observed in the following section. Overall, each of the variables with the exception of Gender appear to be strong predictors of whether or not an individual donated to our charity.

Region (REG)

Region is a categorical variable with four levels: REG1, REG2, REG3, and REG4, which shows the geographic region in which the observed individual resides. In this dataset, this variable has been divided into four (4) dummy variables to represent the five (5) possible regions where the observed individual resides. This variable takes on a value of one (1) if the individual resides in the given region, and a 0 if not. If all four dummy variables are 0s, it is assumed that the respondent lives in the fifth region not represented by a dummy variable. The table below shows the number of individuals by region along with the percentage of individuals in that region who donated to our charity. In Regions 1 and 2, over half of all sampled individuals donated to our charity which could be a good indicator that our marketing efforts should be more focused on these areas. This theory will be teased out during the modeling process.

reg1	Count	App%	donr%
0	3168	79.52%	49%
1	816	20.48%	56%
Total	**3984**	**100.00%**	**50%**

reg2	Count	App%	donr%
0	2645	66.39%	41%
1	1339	33.61%	67%
Total	**3984**	**100.00%**	**50%**

reg3	Count	App%	donr%
0	3492	87.65%	52%
1	492	12.35%	36%
Total	**3984**	**100.00%**	**50%**

reg4	Count	App%	donr%
0	3447	86.52%	53%
1	537	13.48%	34%
Total	**3984**	**100.00%**	**50%**

Homeowner Status (HOME)

HOME is a binary variable that takes on a value of 1 if the observed individual owns a home and 0 if the individual does not own a home. According to the table below, a majority of the observed individuals (88%) are homeowners. 55% of homeowners have donated to the charity, while only 10% of non-homeowners have donated. This variable looks to be another example that can be used in the predictive models based on these initial splits.

home	Count	App%	donr%
0	465	11.67%	10%
1	3519	88.33%	55%
Total	**3984**	**100.00%**	**50%**

child	Count	App%	donr%
0	1395	35.02%	86%
1	404	10.14%	50%
2	1151	28.89%	33%
3	656	16.47%	25%
4	281	7.05%	16%
5	97	2.43%	6%
Total	**3984**	**100.00%**	**50%**

Number of Children (CHLD)

Number of Children is a categorical variable that takes on an integer value which represents the number of children the observed individual has. According to the table below, a person becomes less likely to be a donor as they have more children. 86% of individuals with no children donated to the charity. Based on the DONR% values, it could make sense to create variable simply indicating whether or not the observed individual has a child. Dividing the observations into bins (i.e. 0, 1-2, 3+) could also be a way to consolidate this variable.

Household Income (HINC)

Household Income is a categorical variable that divides observed individuals into seven (7) groups based on their household income. According to the table below, individuals falling into groups 3-5 are more likely to donate than individuals in the other groups. Half of individuals in group 3 donated to the charity, while two-thirds of individuals in group 4 donated.

hinc	Count	App%	donr%
1	212	5.32%	14%
2	467	11.72%	31%
3	407	10.22%	50%
4	1835	46.06%	67%
5	583	14.63%	47%
6	260	6.53%	27%
7	220	5.52%	21%
Total	**3984**	**100.00%**	**50%**

genf	Count	App%	donr%
0	1574	39.51%	51%
1	2410	60.49%	49%
Total	**3984**	**100.00%**	**50%**

Gender (GENF)

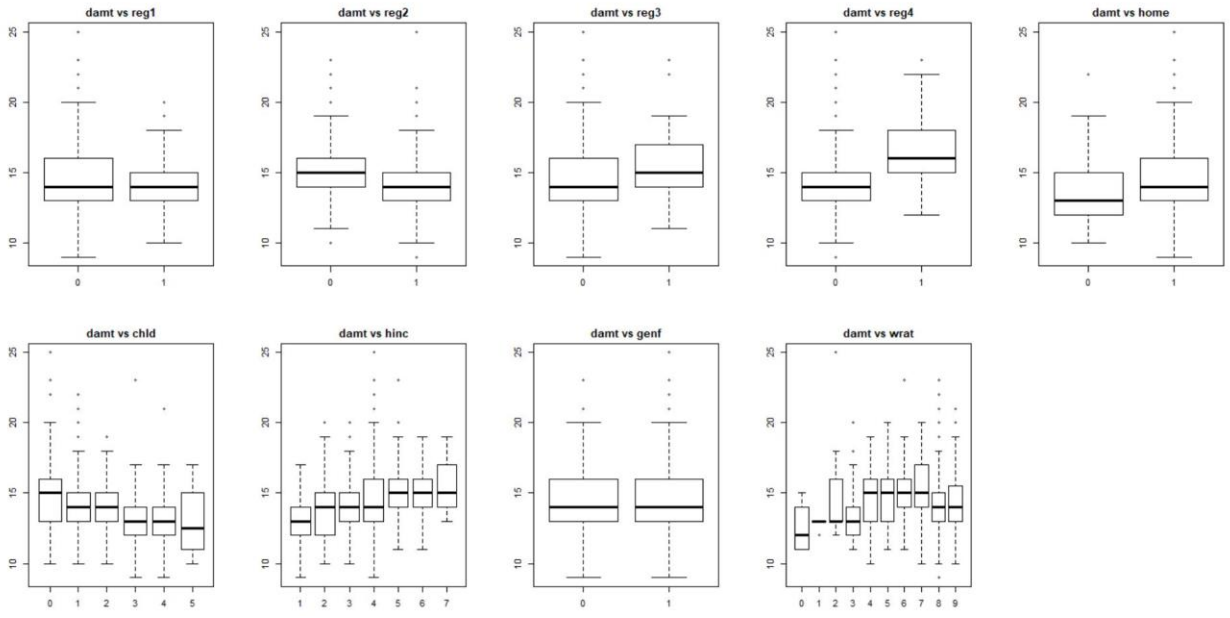
Gender is a binary variable that assigns a value of one (1) if the observed individual is a female, and a value of 0 if the individual is a male. Based on the table below, there does not look to be a substantial difference in giving between males and females. This knowledge could lead us to exclude this variable from final modeling, as it may solely add noise to the model.

Wealth Rating (WRAT)

Wealth Rating is a categorical variable that divides observed individuals into 10 classes based on relative wealth compared to the median income family in the area where they live. Based on this variable, individuals with higher wealth ratings are more likely to have donated to our charity in the past and could serve as ideal targets for our marketing campaigns. Conversely, individuals with lower wealth ratings, particularly ratings of 0 or 1, are much less likely to donate to our charity.

wrat	Count	App%	donr%
0	97	2.43%	5%
1	88	2.21%	6%
2	97	2.43%	27%
3	141	3.54%	22%
4	208	5.22%	34%
5	192	4.82%	41%
6	271	6.80%	55%
7	238	5.97%	59%
8	1557	39.08%	57%
9	1095	27.48%	55%
Total	**3984**	**100.00%**	**50%**

Boxplot for Donation Amount (Responded)

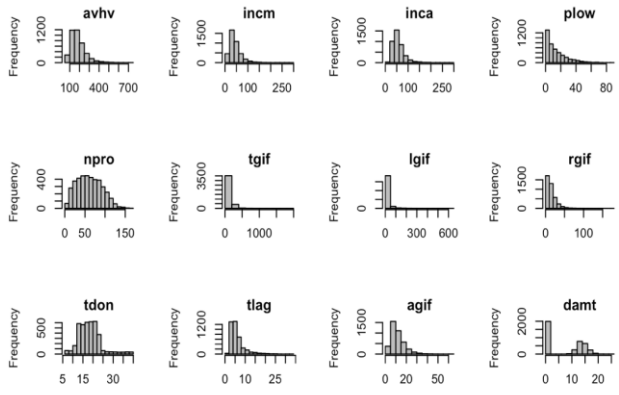
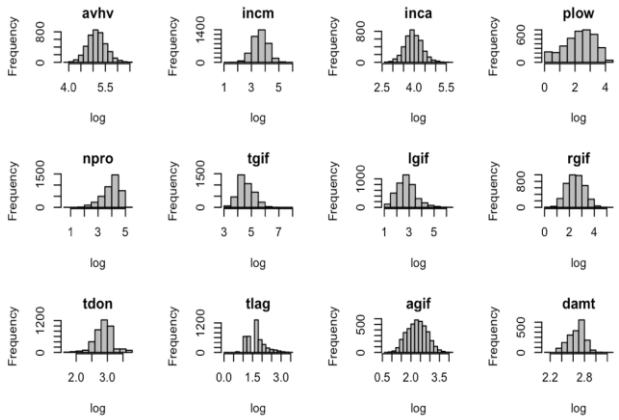


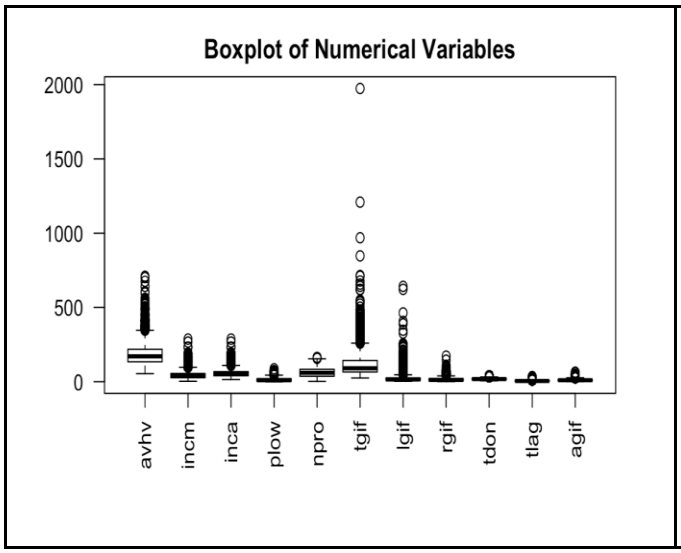
For donation amount, except gender, all the other category attributes have different donation amount distributions.

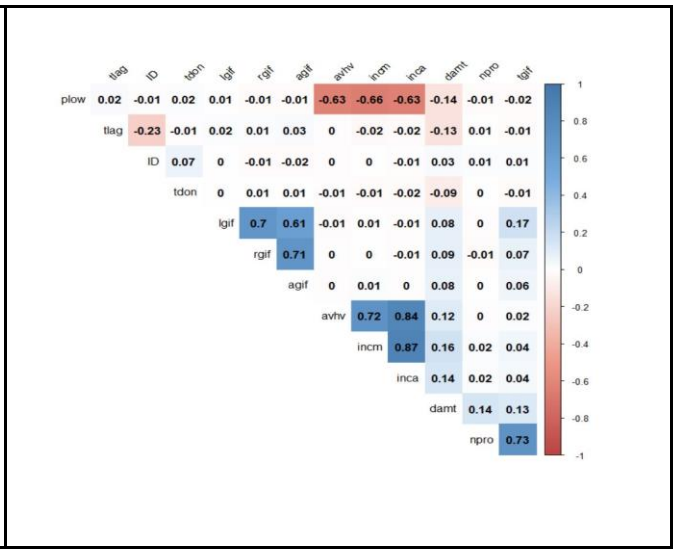
Numeric Predictor Variable Exploration

Among the 17 predictor variables included within the dataset, 12 of the included variables are numeric in nature. There are no missing values of concern across any of these variables. Based on the distributions, we need to trim avhv, incm, inca, plow, npro, tgif, lgif, rgit, and agif.

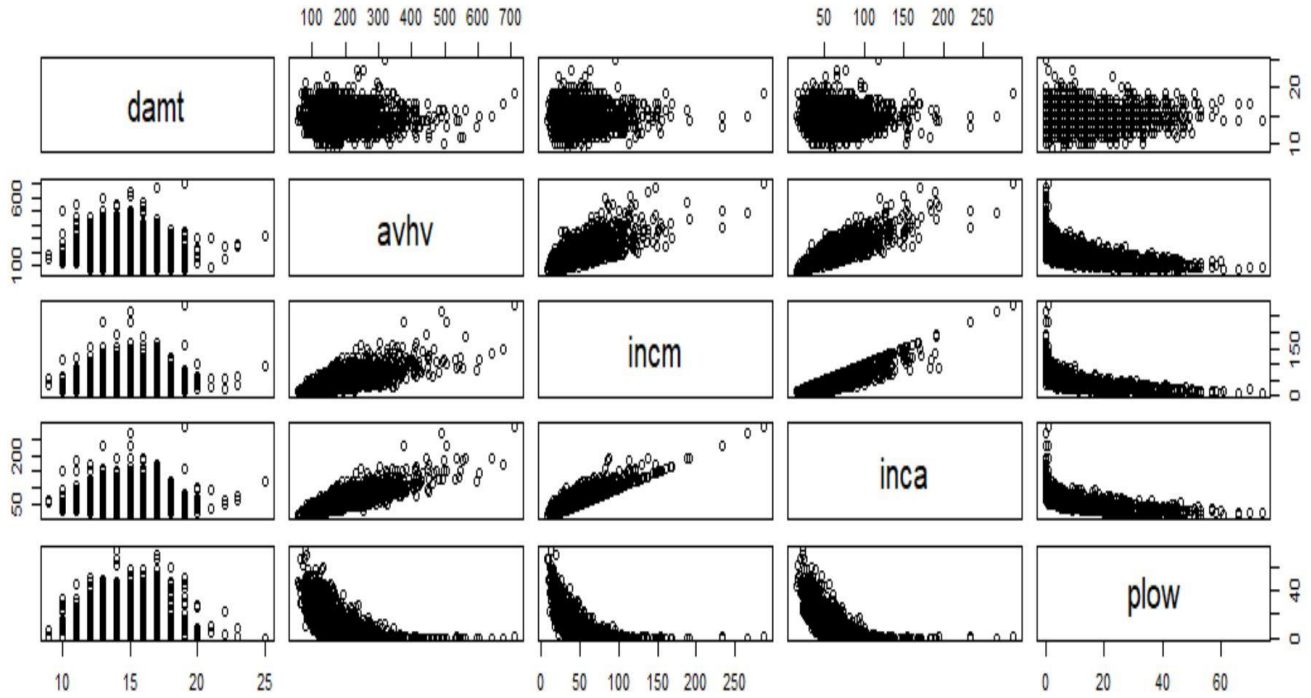
	min	1%	5%	25%	Mean	Median	75%	95%	99%	Max	sd	Zeros	missing
avhv	54	76	94	134	185.18	171	219	324.85	436.85	710	74.7	0	0
incm	3	11	17	27	44.29	39	55	90	131.17	287	25.2	0	0
inca	15	21	28	40	57.14	52	68	103	149	287	25.3	0	0
plow	0	0	0	4	13.73	10	20	41	56	87	13.1	338	0
npro	2	9	16	37	61.63	60	84	113	132	164	30.3	0	0
tgif	25	31	43	65	116.74	91	143	265	402.17	1974	85.5	0	0
lgif	3	4	5	10	23.19	15	25	65.85	156.17	642	31.3	0	0
rgif	1	3	4	7	15.55	12	20	38	57	173	12.1	0	0
tdon	6	8	12	15	18.81	18	22	30	38	40	5.6	0	0
tlag	1	3	3	4	6.3	5	7	13.85	21	34	3.6	0	0
agif	1.9	3.44	4.44	6.99	11.66	10.22	14.79	23.8	33.25	64.22	6.4	0	0

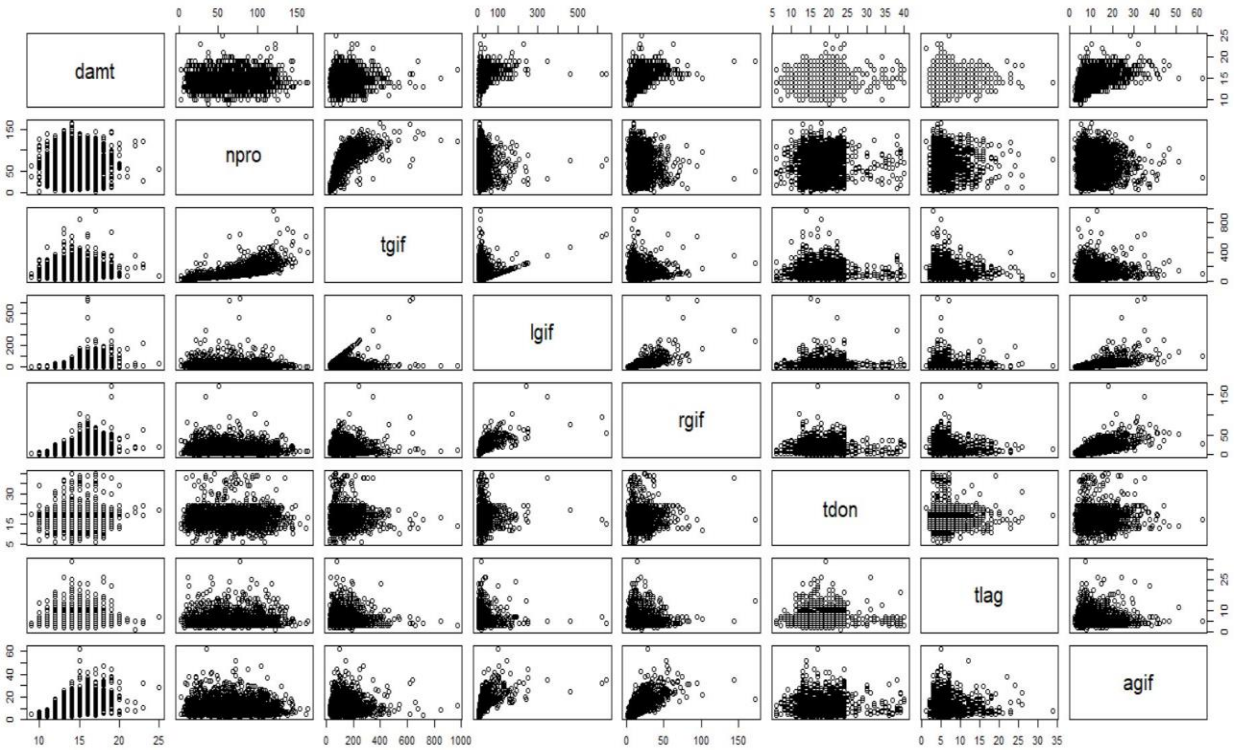
Non-Transformed	Log-Transformed
	





Scatter Plots





Variables Relating to Donor's Neighborhood (AVHV, INCM, INCA, PLOW)

AVHV, INCM, INCA, and PLOW each relate to the observed individual's neighborhood. These variables have been grouped together for the purpose of this EDA, as each demonstrate strong correlation with each other (shown in correlation plot above). With this being said, it will most likely make sense to include just one (1) of these variables to avoid multicollinearity. Prior to any transformations, these variables are each strongly right-skewed due to numerous outliers which are shown in the boxplot. Knowing the extent of these outliers, we may truncate the observations before modeling. Additionally, we will create new, transformed variables to include in our model in order to satisfy the normality assumption in our linear models. Other key metrics for each of these variables are included in the above table.

Lifetime Number of Promotions Received to Date (NPRO)

NPRO measures the number of lifetime promotions received by the observed individual. In theory, we would expect individuals who have been exposed to our charity's marketing to be more likely to donate. NPRO ranges from 2 to 164 with a mean value of 61.63. Other key metrics, such as various percentiles, are included in the table above. This variable looks to be more uniformly distributed with a slight right-skewness. There are a few outliers that appear to be right on the cusp of the upper tail on the boxplot. We will truncate the upper 99% of this variable and proceed with modeling. NPRO appears to have a strong positive correlation with TGIF which will need to be considered when modeling.

Variables Relating to Gifts/Donations (TGIF, LGIF, RGIF, TDON, TLAG, AGIF)

There are numerous variables related to past giving that will be examined in this portion of the EDA. Each of the giving variables have a noticeable right-skewness which appears to be a function of a few outliers on the upper end of the distribution. We will effectively eliminate some of these outliers by truncating the upper 99% of the data in order to eliminate the possibility of these outliers negatively impacting our models. The log-transformed version of each of these variables are much more normally distributed which will lead us to ultimately use the log-form of these variables when modeling. AGIF, LGIF, and RGIF each are strongly correlated which we will need to be cautious of when modeling. When building the models, we will check VIF values to ensure multicollinearity is not of concern.

Summary for Attributes Transformation for Modeling

#	Attributes will be used for modeling	Original Name	Description	Cap (For Original Attributes)
1	reg1	reg1	Indicate whether the potential donor belong to Region 1	
2	reg2	reg2	Indicate whether the potential donor belong to Region 2	
3	reg3	reg3	Indicate whether the potential donor belong to Region 3	
4	reg4	reg4	Indicate whether the potential donor belong to Region 4	
5	home	home	1= homowner, 0 = not a homeowner	
6	chld	chld	Number of children	
7	chld_0	child	Whether the donar do not have child	
8	chld_ThM	child	Whether the donar has children >=3	
9	hinc	hinc	Household income (7 categories)	
10	hinc_4	hinc	Whether household income category is 4	
11	genf	genf	Gender (0=Male, 1= Female)	
12	wrat	wrat	Wealth Rating	
13	wrat_0_4	wrat	Whether wealth rating is between 0 and 4	
14	avhv_log	avhv	avg home value $k	450
15	incm_log	incm	Median family income	130
16	inca_log	inca	avg family income	150
17	plow	plow	% low income	56
18	npro	npro	# promotion recveived lifetime	132
19	tgif	tgif	$ gifts lifetime	402
20	lgif	lgif	max $ gifts	156
21	tdon	tdon	# months since last donation	
22	tlag	tlag	#months between 1st & 2nd gift	
23	rgif_log	rgif	$ gift most recent	60
24	agif_log	agif	avg $ gift	35
25	GiftTimes	tgif & agif	total gifts divided by avg should be the total reposed times	
26	ResponseRate	tgif, agif, npro	GiftTimes divided by the toal promotion times (npro)	

New attributes
Transformed -log

Analysis

Classification Models

The first types of models that will be developed are classification models which will attempt to predict whether or not a chosen individual will donate to our charity. We have standardized all the potential predictors to be easily applied to different algorithms.

Model	Model- R code	# Donor	*Precision	Error Rate	Maximum Profit	
Boosting	boost.model	1240	80.2%	8.77%	$ 11,947.5	Best
Logistic Regression GAM	gam.lr	1291	76.9%	15.06%	$ 11,816.5	
Logistic Regression	glm.stepwise	1255	78.7%	13.78%	$ 11,816.0	
LDA	model.lda	1279	77.3%	14.87%	$ 11,782.5	
QDA	model.qda	1229	77.8%	15.66%	$ 11,404.0	
SVM	SVM.best	1060	87.5%	10.00%	$ 11,336.0	
Random Forest	RF.Model	1058	87.2%	10.46%	$ 11,267.5	
KNN	post.valid.knn	1219	76.3%	17.84%	$ 11,018.0	
Decision Tree	TreeModel	1106	80.1%	16.35%	$ 10,649.5	

** Precision =TP/(TP+FP)*

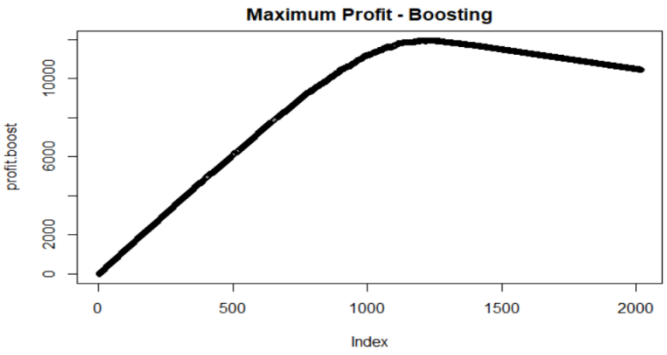
Above is the list of candidate models for donor classification. We built the builds on train data and test on valid data set. Based on maximum profit, the best classification model is "Boosting". We achieve $11,947.5 for boosting model. For the valid data set, we have 2018 observations. The boosting classification model classifies 1240 observations as donors. The true donor takes 80.2% of those 1240. Overall, the model error rate is 8.77%. Compare with other model, boosting has an outstanding error rate. The second best model is Logistic Regression GAM, the maximum profit is $11,816.5, which is very close to boosting ($131 less). But the error rate is almost doubled (15% compared to 8.77%).

Boosting Classification Model

We have ranking the individuals in the valid data set by boosting predictive score. The higher the score, means the individual is more likely to donate based on boosting model. If it is a true donor, then we will get 14.5 profits. All the mailed individuals will cost us 2 dollars. From the "Maximum Profit –Boosting" graph, it shows that the maximum achieve is close to $12000. When the scores gets lower, the individual more likely not response, the cost is more than the added profit. The maximum profit achieved when we mailed to 1240 individuals out of the 2018 valid data set to get $11947.5 profit.

Confusion Matrix

		Valid	
Boosting		0	1
	0	774	4
	1	245	995



This is the boosting classification n

```
boost.model < - gbm(donr ~ reg1 + reg2 + home + chld + wrat +
                    npro + tgif + tdon + tlag + chld_0 + hinc_4 + wrat_0_4 +
                    incm_log + I(hinc^2) + npro * tgif,
```

```
data = data.train.std.c, distribution = "bernoulli",
shrinkage = 0.1, n.minobsinnode = 10, n.trees = 150, interaction.depth = 3)
```

Predictios used in the boosting classification model: reg1, reg2, home, chld, wart, npro, tgif, tdon, tlag, chld_0, hinc_4, wart_0_4, incm_log, I(hinc^2) and npro * tgif.

Logistic Regression –GAM

Below is the best logistic regression –GAM, which predict 1291 donars from the 2018 valid data set. The maximum profit achieve is 11816.5.

```
Call: gam(formula = donr ~ reg1 + reg2 + home + chld + hinc + wrat +
    plow + npro + tgif + tdon + tlag + chld_0 + hinc_4 + wrat_0_4 +
    avhv_log + incm_log + I(hinc^2) + npro * tgif, family = "gaussian",
    data = data.train.std.c)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.97172 -0.22095  0.02156  0.22939  0.79906

(Dispersion Parameter for gaussian family taken to be 0.0982)

    Null Deviance: 995.9977 on 3983 degrees of freedom
Residual Deviance: 389.3497 on 3965 degrees of freedom
AIC: 2081.057

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
             Df Sum Sq Mean Sq   F value    Pr(>F)
reg1          1   3.17   3.175   32.3283 1.396e-08 ***
reg2          1  81.94  81.935  834.3972 < 2.2e-16 ***
home          1  77.87  77.870  792.9984 < 2.2e-16 ***
chld          1 258.16 258.164 2629.0526 < 2.2e-16 ***
hinc          1   0.55   0.551    5.6098  0.017908 *
wrat          1  32.74  32.743  333.4466 < 2.2e-16 ***
plow          1  13.22  13.223  134.6586 < 2.2e-16 ***
npro          1  12.01  12.015  122.3541 < 2.2e-16 ***
tgif          1   1.56   1.561   15.8975 6.807e-05 ***
tdon          1   3.20   3.203   32.6204 1.202e-08 ***
tlag          1  10.22  10.217  104.0432 < 2.2e-16 ***
chld_0        1  27.26  27.262  277.6228 < 2.2e-16 ***
hinc_4        1  64.37  64.370  655.5183 < 2.2e-16 ***
wrat_0_4      1   1.90   1.903   19.3791 1.100e-05 ***
avhv_log      1   0.69   0.690    7.0223  0.008082 **
incm_log      1   3.30   3.301   33.6140 7.244e-09 ***
I(hinc^2)     1  13.78  13.785  140.3779 < 2.2e-16 ***
npro:tgif     1   0.68   0.682    6.9478  0.008425 **
Residuals  3965 389.35   0.098
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

[1]  1291.0 11816.5
                 c.valid
chat.valid.gam.lr   0   1
                0 721   6
                1 298 993
```

Prediction Models

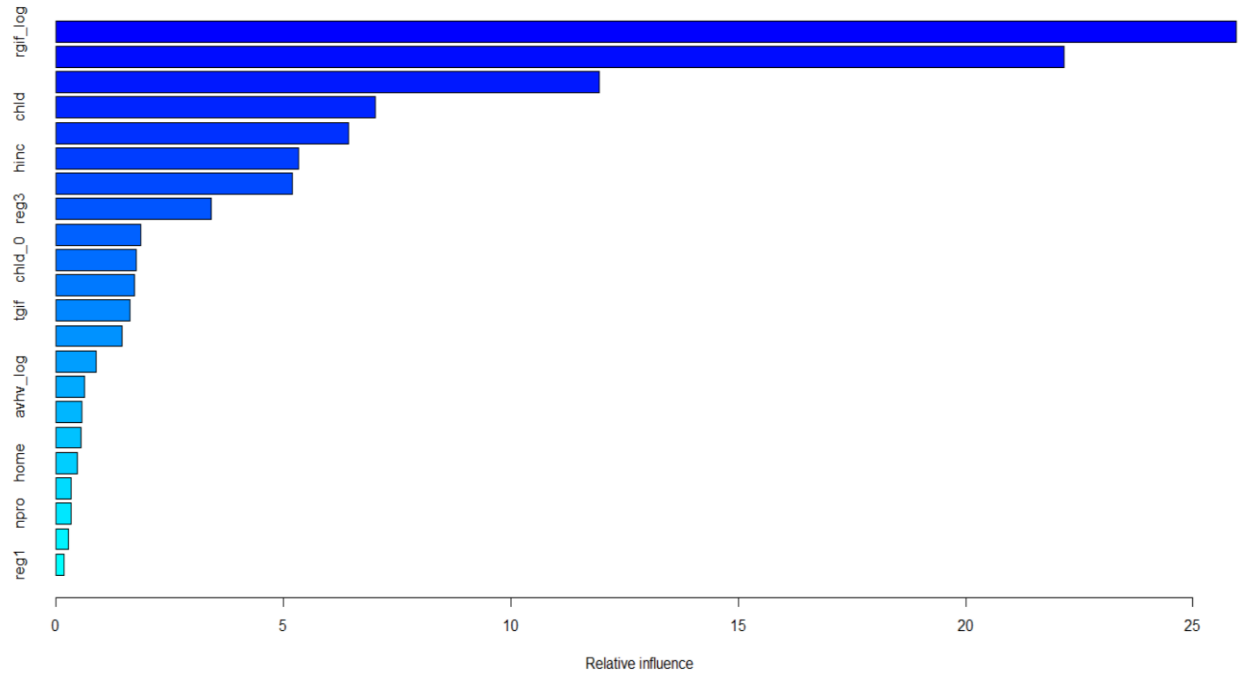
Now as we determined the best classification model, we will focus on developing predictive models to predict donation amounts based on donor features. We have developed nine different types of predictive models (Boosting, Least Squares Regression with all the attributes, Best Subset Selection with BIC, Best Subset Selection with 10 K-Fold Cross-Validation, Principle Components Regression, and Ridge Regression, Lasso Regression, Partial Lease Squares, and

Random Forests) . Based on the model with the lowest mean prediction error in the valid data set, Boosting is also the best model for predicting donation amounts. It has slight lower mean prediction error compared with the other 8 models. The boosting model has mean prediction error 1.530216. The standard error is 0.1613 for the valid data set, which is higher than the Least Square with all the attributes. Since we use mean predictor error as the measurement for model selection, we will still use boosting model. However, for more stable performance and simple interpretation, I prefer "Best subset-BIC" or the "OLS_Full".

#	Model	mean prediction error	sd
1	Boosting	1.530216	0.1613
2	OLS_Full	1.542719	0.1557
3	Best subset -BIC	1.543904	0.1555
4	lasso	1.581835	0.1556
5	PCA	1.597288	0.1551
6	ridge	1.60079	0.1582
7	Partial Least Squares	1.612741	0.1543
8	10 CV	1.614235	0.1582
9	Random Forest	1.721304	0.1689

Boosting Prediction Model

The graph below shows the variable importance for boosting regression. rgif_log , agif_log, reg4, chld, lgif, hinc, wrat, and reg3 have much higher importance in predicting donation amount in the boost model.

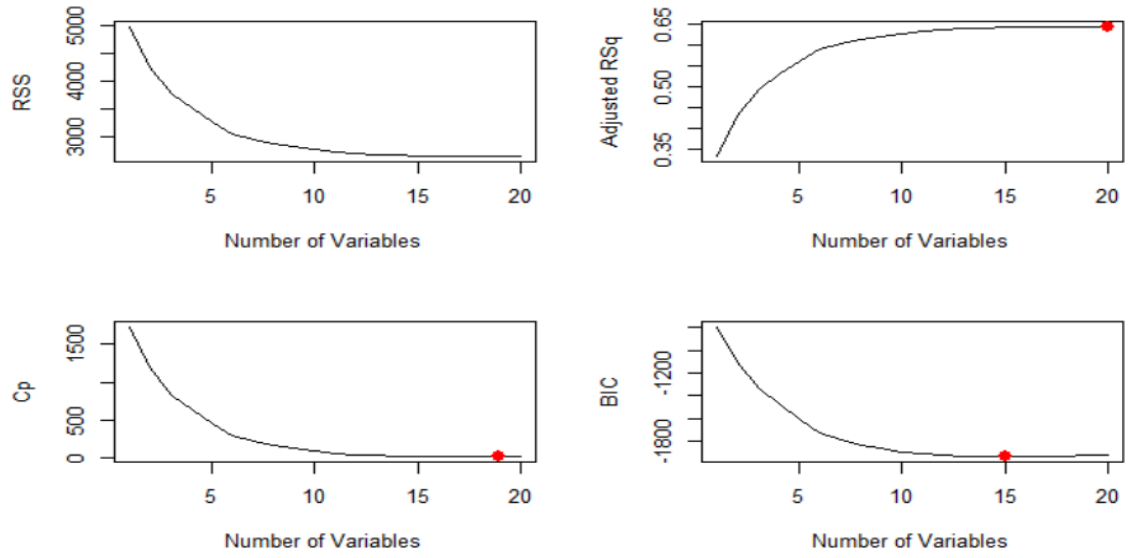


Below is the boosting regression model:

```
boost.pred <- gbm(damt ~rgif_log+agif_log+reg4+chld+lgif+hinc+wrat+reg3+
                  chld_0+incm_log+plow+reg2+tgif+chld_ThM+inca_log+avhv_log+
                  home+ResponseRate+npro+tdon+reg1+GiftTimes,
                  data = data.train.std.y, distribution = "gaussian",
                  n.trees=5000, interaction.depth=5)
```

Best subset – BIC/CP/Adjusted R^2

For the best subset with adjust R^2, the number of parameters selected is 20. With Mallow's Cp, it minimized with 18 variables. For BIC, it minimized with 15 variables. Based on those graphics, adding additional variables, the performance improvement is not significant after 15. I will choose 15 for the best subset based on BIC, Cp, and Adjusted R^2.



Below is the best subset model with BIC:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.07522    0.04914 286.443  < 2e-16 ***
reg3           0.33710    0.03058  11.022  < 2e-16 ***
reg4           0.65515    0.03141  20.855  < 2e-16 ***
home           0.25520    0.05514   4.628 3.93e-06 ***
chld          -0.66540    0.03432 -19.388  < 2e-16 ***
hinc           0.50455    0.03618  13.944  < 2e-16 ***
wrat          -0.35908    0.05732  -6.264 4.59e-10 ***
plow           0.21522    0.05819   3.699 0.000223 ***
tgif           0.28540    0.03473   8.218 3.69e-16 ***
wrat_0_4      -0.51450    0.05749  -8.950  < 2e-16 ***
incm_log       0.31947    0.05096   6.269 4.45e-10 ***
rgif_log       0.65321    0.04402  14.837  < 2e-16 ***
agif_log       0.55733    0.04405  12.651  < 2e-16 ***
I(hinc^2)     -0.07553    0.02651  -2.849 0.004438 **
plow:inca_log -0.15476    0.02934  -5.274 1.48e-07 ***
tgif:npro     -0.09560    0.02633  -3.631 0.000289 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.159 on 1979 degrees of freedom
Multiple R-squared:  0.6446,    Adjusted R-squared:  0.6419
F-statistic: 239.3 on 15 and 1979 DF,  p-value: < 2.2e-16
```

Results

We get the best result with boosting for both classification and prediction model. For the donor classification, the boosting model achieve maximum profit $11.947.5. For the prediction model, it has 1.530216 mean prediction error, and 0.1613 standard error.

After we applied the boosting classification model, and boosting prediction models to the test data set, we get the response rate 15.05%. Compared to our recent mailing records, the typical overall response rate is 10%. The boosting model will help us increase response rate by 50.5%. Compared with the average donation amount, the boosting model get $14.10, which is lower than $14.50 for our donor who response to the mailing. Because we do not know what is the response rate for the test data set. The average donation amount is based on all the test data set. If we only average the yhat for the chat equals to 1, we get average donation amount $14.42, which is extremely close to $14.50. This means both the boosting classification model and prediction model works quite well.

```
     chat             yhat
Min.   :0.0000   Min.   :10.64
1st Qu.:0.0000   1st Qu.:13.10
Median :0.0000   Median :14.03
Mean   :0.1505   Mean   :14.10
3rd Qu.:0.0000   3rd Qu.:15.11
Max.   :1.0000   Max.   :18.04
```

```
     chat        yhat
Min.   :1   Min.   :11.21
1st Qu.:1   1st Qu.:13.36
Median :1   Median :14.34
Mean   :1   Mean   :14.42
3rd Qu.:1   3rd Qu.:15.52
Max.   :1   Max.   :18.04
```

Based on current maling plan:

Expected profit: $14.50*0.10-2 =-$0.55

Based on Boosting model:

Expected profit: $14.42*0.1505-2 =-$0.55 =$0.17021

In the past, it is will be a loss for us if we mailed to everyone. But if we use the boosting models, we will have positive profit. If we have to be more aggressive, we can increase the thresholds to only promote to the ones more highly likely to response.

Conclusion

We have successfully building the classification model with boosting, which will increase our mailing response rate by 50.5% (from 10% to 15.05%). The boosting prediction model is also very accurate. For the predicted responded individuals, the average donation amount is $14.42, which is very close to $14.50. By implementing those models, expect to get $0.17021 profit from each mail. We can only select the individuals have the highest scores to market to if we want higher profit per mailer.