Noé Flores
Charity Project:

Introduction:
The problem presented for us is to assist a charitable organization that is looking to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to previous donors. Direct marketing campaigns require a great deal of analysis in order to effectively manage the costs associated with this particular method of targeting potential donors most likely to contribute again. Our goal is to develop a machine learning model to determine the likely chance of a response from a donor and estimate the dollar amount of the predicted donation. We will begin by performing an exploratory analysis of our data set to gather information regarding the variables we have available to incorporate into this project and follow up with any data preparation needed to ensure that we are working with optimal data. Finally, we will build and compare several models to determine which models have the most significant predictive value, interpretability, and practicality with respect to our goal of identifying likely donors and donation amounts. The results of this analysis should produce models we can present to the charity which will improve the effectiveness of their campaign.

Data:
Our dataset contains 8,009 observations of 22 different explanatory variables and 2 response variables. The table below provides a breakdown and description of those variables.

Figure 1: Variables and Descriptions.

Independent Variables	Description
REG1 REG2 REG3 REG4	Region which the donor belongs to (1 = belongs to region 0 = does not belong)
HOME	(1 = homeowner 0 = not a homeowner)
CHLD	Number of children
HINC	Household income (7 categories)
GENF	Gender (0 = Male 1 = Female)
WRAT	Wealth Rating (0-9 with 0 being lowest wealth)
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
NPRO	Lifetime number of promotions received to date
TGIF	Dollar amount of lifetime gifts to date
LGIF	Dollar amount of largest gift to date
RGIF	Dollar amount of most recent gift
TDON	Number of months since last donation
TLAG	Number of months between first and second gift
AGIF	Average dollar amount of gifts to date
TLAG	Number of months between first and second gift
AGIF	Average dollar amount of gifts to date
Dependent Variables	**Description**
DONR	Classification Response Variable (1 = Donor, 0 = Non-donor)
DAMT	Prediction Response Variable (Donation Amount in $)

Exploratory Data Analysis:
Once the variables are converted, we can examine their measures of central tendency. We will focus on the integer variables since the other variables represent more categorical styled data. In some instances, there appears to be a relatively broad range which could be an indication of potential skewness and outliers.

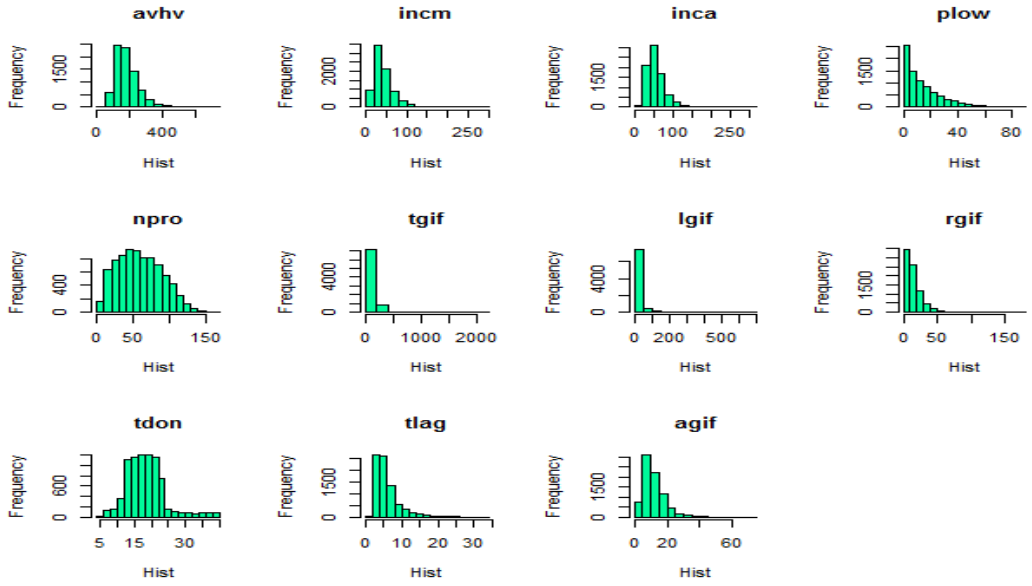
Figure 2: Measures of Central Tendency

avhv	incm	inca	plow	npro	tgif
Min. : 48.0	Min. : 3.00	Min. : 12.00	Min. : 0.00	Min. : 2.00	Min. : 23.0
1st Qu.:133.0	1st Qu.: 27.00	1st Qu.: 40.00	1st Qu.: 4.00	1st Qu.: 36.00	1st Qu.: 63.0
Median :169.0	Median : 38.00	Median : 51.00	Median :10.00	Median : 58.00	Median : 89.0
Mean :182.6	Mean : 43.47	Mean : 56.43	Mean :14.23	Mean : 60.03	Mean : 113.1
3rd Qu.:217.0	3rd Qu.: 54.00	3rd Qu.: 68.00	3rd Qu.:21.00	3rd Qu.: 82.00	3rd Qu.: 137.0
Max. :710.0	Max. :287.00	Max. :305.00	Max. :87.00	Max. :164.00	Max. :2057.0

lgif	rgif	tdon	tlag	agif
Min. : 3.00	Min. : 1.00	Min. : 5.00	Min. : 1.000	Min. : 1.29
1st Qu.: 10.00	1st Qu.: 7.00	1st Qu.:15.00	1st Qu.: 4.000	1st Qu.: 6.97
Median : 16.00	Median : 12.00	Median :18.00	Median : 5.000	Median :10.23
Mean : 22.94	Mean : 15.66	Mean :18.86	Mean : 6.363	Mean :11.68
3rd Qu.: 25.00	3rd Qu.: 20.00	3rd Qu.:22.00	3rd Qu.: 7.000	3rd Qu.:14.80
Max. :681.00	Max. :173.00	Max. :40.00	Max. :34.000	Max. :72.27

The histograms of the numeric variables give us a better picture of the skewness in our data.

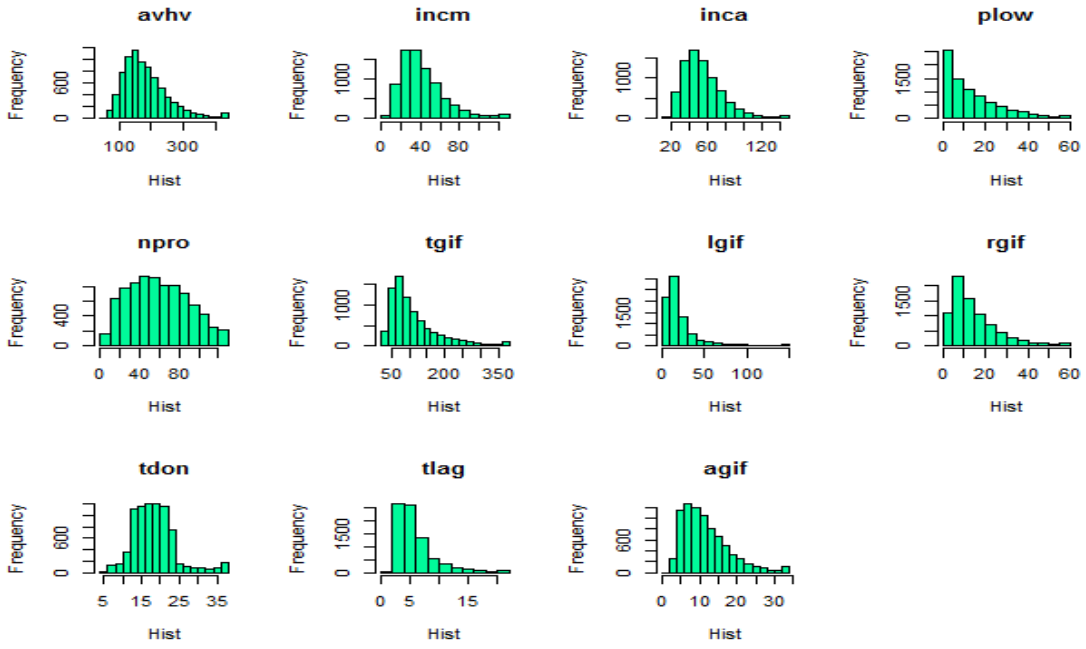
Figure 3: Histogram of numeric variables: Original State.



It's clear from the histograms above that we have some variables which are showing a good amount of positive skewness which means the right tail is longer and the mass of the data is concentrated on the left. The heavy left sided data concentration could be indicative of outliers, but the most likely case is that our dataset is loaded near zero.

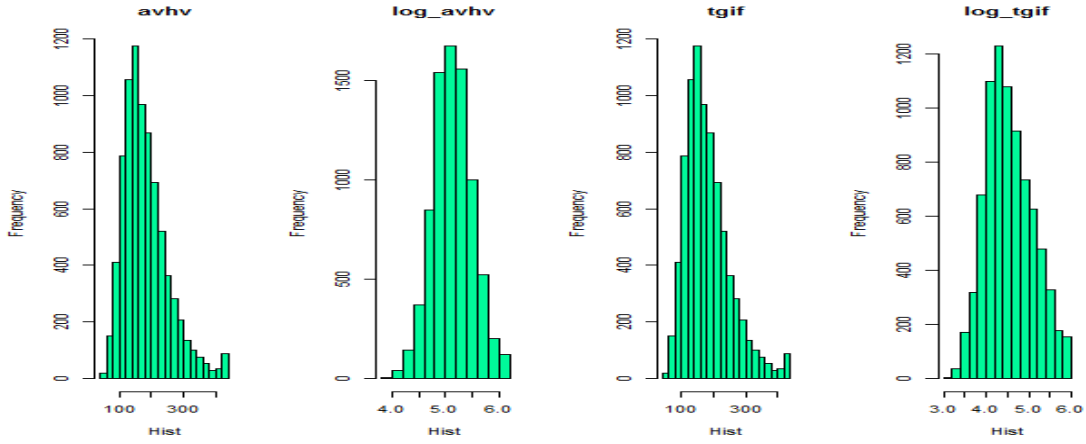
Examination of the box-plots for our variables appears to confirm the potential for outlying data points. Outlying data can pose significant problems for models, but there is also danger in removing data that has been collected and recorded correctly. Data located outside acceptable bounds might be there to point us to potential difficulties of trying to fit a model with the particular approach we selected. This could lead to overfitting. In a conservative effort to finding the middle ground, we will cap the values at the 99% quantile. Below are the histograms representing those changes.

Figure 4: Histogram of numeric variables maxed at 99% quantile.



The histograms above show an improvement over the original data points, but we still see some of that right-sided skewness. A final adjustment will be made in the form of variable transformations for some of our variables. We will log transform AVHV, TGIF, and LGIF. There could be an argument to transform other variables as well, but there is a danger with log transformations in that you may hinder the interpretation of our results. Below is a side comparison of one variable's log transformation compared to its original state.

Figure 5: Normal_Hist vs. Log_Hist.



In terms of AVHV(average home value), you can see that there is an improvement in the distribution and it now looks more normally distributed.In order to get a sense of the relationship between our dependent variables (DAMT and DONR) and our independent variables, we constructed a correlation table for each one. The results below don't appear to indicate anything extraordinary or pronounced with respect to correlation. The findings could also be indicative of the need for a model outside the realm of linearity.

Figure 6: Correlation Table.

DAMT	Corr	DONR	Corr
reg1	0.035	reg1	0.043
reg2	0.214	reg2	0.253
reg3	-0.086	reg3	-0.106
reg4	-0.077	reg4	-0.119
home	0.290	home	0.292
chld	-0.553	chld	-0.533
hinc	0.051	hinc	0.035
genf	-0.008	genf	-0.009
wrat	0.231	wrat	0.238
avhv	0.115	avhv	0.116
incm	0.145	incm	0.139
inca	0.129	inca	0.126
plow	-0.125	plow	-0.132
npro	0.146	npro	0.140
tgif	0.126	tgif	0.115
lgif	0.077	lgif	0.018
rgif	0.078	rgif	0.008
tdon	-0.092	tdon	-0.096
tlag	-0.124	tlag	-0.130
agif	0.078	agif	0.005
donr	0.982	damt	0.982

We can also make use of some of the categorical variables to produce frequency plots to get a better understanding of the relationship between the predictor variables and those who have donated in the past. For reference, a zero (0) value indicates that the variable is negative for that category, meaning they are not a homeowner. A one (1) means the variable is true and the person is a homeowner.

Figure 7: Donor vs. Homeowner

		Donor	
		No	Yes
Homeowner	0	417	48
	1	1572	1947

- There were a total of 1,995 donors, of those donors 1,947 were also homeowners.

Figure 8: Donor vs. Number of Children.

		Donor	
		No	Yes
# of Children	0	194	1201
	1	203	201
	2	770	381
	3	494	162
	4	237	44
	5	91	6

- We can see in this table that as the number of children increases, the total number of donors begins to decrease. Those individuals without children were the most likely to donate.

Figure 9: Donor vs. Household income.

		Donor	
		No	Yes
Household Income	1	183	29
	2	322	145
	3	202	205
	4	609	1226
	5	309	274
	6	190	70
	7	174	46

- Household income (HINC) has 7 different categories starting from lowest to highest. An exciting trend here as we can see donors increasing as household income increases but only up until level 4. After HINC reaches level 4, we begin to see a drop in donors.

Figure 10: Donor vs. Wealth Rating.

		Donor	
		No	Yes
Wealth Rating	0	92	5
	1	83	5
	2	71	26
	3	110	31
	4	138	70
	5	113	79
	6	121	150
	7	97	141
	8	676	881
	9	488	607

- Wealth Rating is calculated using median family income and population statistics from each area to index relative wealth in each state. The regions are denoted from 0 to 9 with 9 the highest wealth category. As you can see from the table, we get a pretty steady increase in donors as we move up wealth ratings, but the donors vs. non-donors is pretty much evenly split.

Figure 11: Donor vs. Gender

		Donor	
		No	Yes
Gender	0	769	805
	1	1220	1190

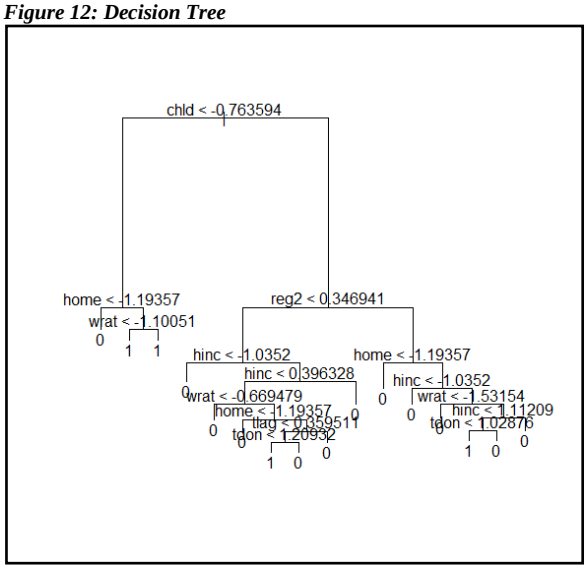
- There were a total of 1,190 female donors and 805 male donors. It would appear from this split that female donors tend to be slightly more inclined to donate.

Modeling: Classification models.
Our work will initially focus on the development of a classification model for donors (DONR). In an effort to be thorough we will fit nine different classification models. We won't devote a lot of time to explanations of the modeling but instead, focus on the results each model produced. Our dataset was broken down into training, testing, and validation subsets. The modeling was performed on the training data and tested on the validated set.

We started our process with a Logistic Model. These models aren't typically built for classification problems, but they represent a powerful tool that is flexible to use. The outcome of the Logistic model actually gave us decent results, but it was clear that the data could benefit from more robust modeling.

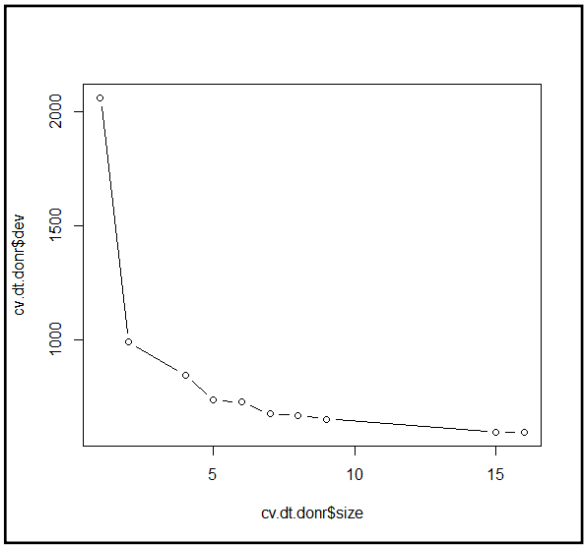
Our attention next turned to Tree-Based regression methods. The general idea is that we segment the predictor space into a number of simple regions and work from there.

Figure 12: Decision Tree



- Here we see the splits in our decision tree. It's clear the primary driver for donations appears the be the number of children. Trees can be pruned through cross-validation to get a better sense of how many predictor variables are actually significant for modeling.

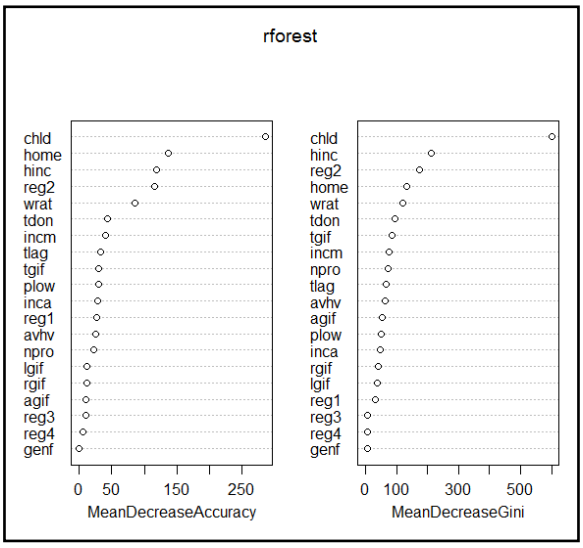
Figure.13:Cross Validation



- The results of the cross-validation appear to indicate that a model could be fit using five variables and at the most ten since the most prominent bend in the plot happens at value = 5. The curve in the graph does continue, so we felt it prudent to build the decision tree model with 9 variables and validated those results.

A subset of Decision Trees are the Random Forest models. We decided to fit the random forest model to see if any improvement could be made on the classification metrics.

Figure 14: Random Forest



- The Random Forest outputs a variable of importance feature. This feature provides a more detailed view of the predictors that drive the model. Once again we see chld, home,hinc, reg2, wrat driving the model.

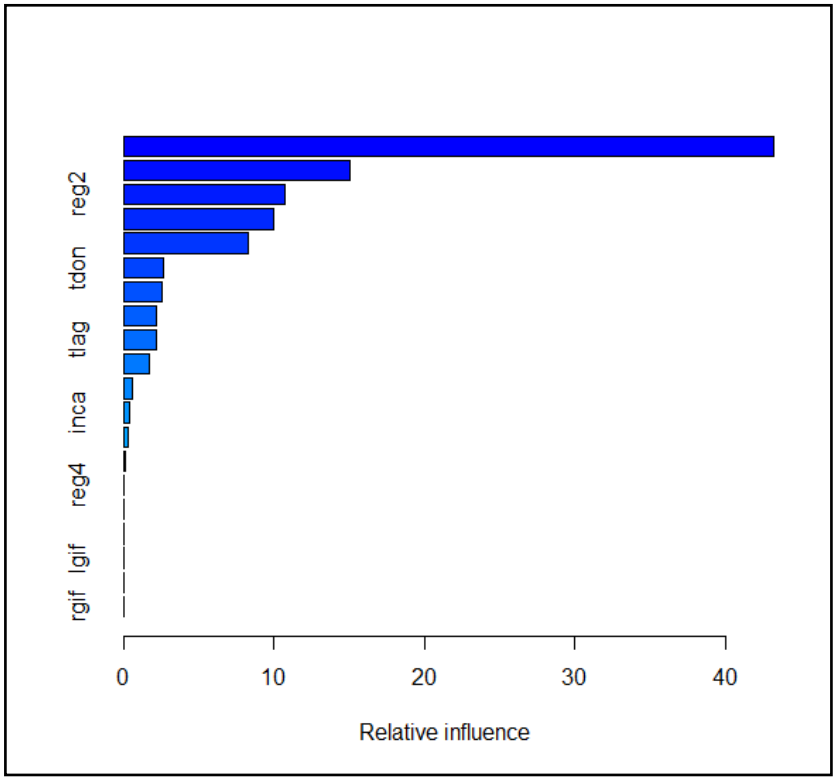
- The Random Forest Gini feature of importance provides us with another measuring of relevance. Interestingly enough, the variables traded places but nothing vastly different.

Boosting is another method we can use to improve upon our Random Forest Models. It has three tuning parameters, The number of trees, selected through cross-validation, the shrinkage parameter, and the number of splits in each tree. The variable significance is shown below.

Figure 15: Boosting

Boosting

var	rel.inf
chld	43.226330717
hinc hinc	15.015753253
reg2 reg2	10.750467749
home home	9.999621188
wrat wrat	8.273751980
tdon tdon	2.639065520
incm incm	2.541707091
tgif tgif	2.211717499
tlag tlag	2.208722962
reg1 reg1	1.714093112
plow plow	0.586092207
inca inca	0.419417384
npro npro	0.255523387
avhv avhv	0.127416591
reg4 reg4	0.012921232
reg3 reg3	0.011944487
agif agif	0.003481152
lgif lgif	0.001972490
genf genf	0.000000000
rgif rgif	0.000000000



The nine models consist of Logistic regression, Logistic regression(GAM), Linear Discriminate Analysis (LDA), K-nearest neighbors(KNN), Decision Tree, Bagging (aggregate decision tree method), Boosting (aggregate decision tree method), and Support Vector Machine (SVM). Each of these modeling techniques allows us to produce a confusion matrix, which is a table that is used to describe the performance of a classification model. Those results are below.

Figure 16: Classification Model Confusion Matrix

		Donor			
		No	Yes		
Logistic Model	0	703	16	**Sensitivity**	68.99%
	1	316	983	**Specificity**	98.40%
		Donor			
		No	Yes		
Logistic GAM	0	575	13	**Sensitivity**	56.43%
	1	444	986	**Specificity**	98.70%
		Donor			
		No	Yes		
LDA	0	680	12	**Sensitivity**	66.73%
	1	339	987	**Specificity**	98.80%
		Donor			
		No	Yes		
QDA	0	699	30	**Sensitivity**	68.60%
	1	320	969	**Specificity**	97.00%
		Donor			
		No	Yes		
Boosting	0	737	7	**Sensitivity**	72.33%
	1	282	992	**Specificity**	**99.30%**
		Donor			
		No	Yes		
KNN	0	721	91	**Sensitivity**	70.76%
	1	298	908	**Specificity**	90.89%
		Donor			
		No	Yes		
Decision Tree	0	882	174	**Sensitivity**	86.56%
	1	137	825	**Specificity**	82.58%
		Donor			
		No	Yes		
Bagging	0	877	76	**Sensitivity**	86.06%
	1	142	923	**Specificity**	92.39%
		Donor			
		No	Yes		
SVM	0	828	133	**Sensitivity**	81.26%
	1	191	866	**Specificity**	86.69%

The Sensitivity and Specificity are statistical measures we can use to assess the performance of our classification models. Our interest is in specificity and if you compare the values across the tables above in Figure.8, you will see that the boosting model performed the best at 99%.

Classification Model Selection:
Our final model selection criteria will be based on maximum predicted profit from each model. The table below provides a comparison of the profit anticipated for each classification model as well as their mean prediction error rate.

Figure 17: Classification Model Predicted Profit Summary

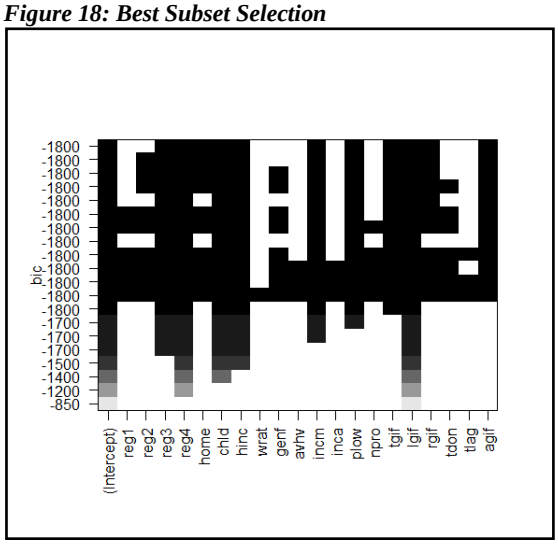
Model	Donors	Profits	Mean Error
Boosting	1274	$ 11,836.00	0.108
LDA	1326	$ 11,659.50	0.131
Logistic Model	1299	$ 11,655.50	0.121
QDA	1289	$ 11,472.50	0.151
Logistic GAM	1430	$ 11,437.00	0.163
Baggimg	1065	$ 11,253.50	0.111
KNN	1206	$ 10,754.00	0.193
SVM	1057	$ 10,443.00	0.161
Decision Tree	962	$ 10,038.50	0.154

The table above is sorted according to predicted profited from largest to smallest predicted value. The model which incorporates boosting is shown to perform best in terms of maximum profit. This model also exhibits the smallest mean error. With that said, it's clear we should continue with the Boosting model for our prediction of possible donor(DONR) response.

Modeling: Prediction Models.
The second task was to develop a model that could help predict the donation amount of potential donors. Much like the system we used for the classification models, we will look to develop nine different regression models to help us predict donation amount (DAMT). The models will consist of different versions of Ordinary Least Squares regression, Best Subset using Forward Selection, Ridge Regression, Lasso Regression, Decision Tree Model, Bagging, Random Forest, and Boosting. As was the case in the classification models, the modeling was performed on the training data and tested on the validated set.

We began our development with a simple linear regression model, but if you recall the earlier exploratory data analysis on page.4, many of the variables we have in our dataset aren't highly correlated with donation amount(DAMT). This would suggest that we need to move towards a more robust selection process. We started that process with best subset selection using a forward selection process.

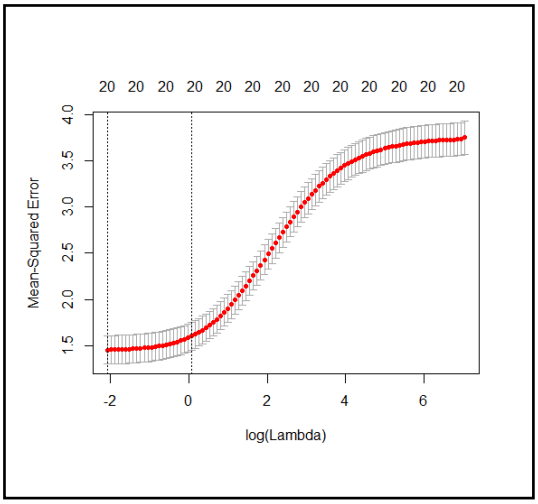
Figure 18: Best Subset Selection



- We used best subset selection with a forward process to select the best model. The best model, in this case, is the one with the lowest BIC. The plot on the left represents those results. Each line in the plot represents a model and the corresponding BIC value.

Ridge Regression is similar to least squares, but it intends to estimate coefficients by minimizing the lambda tuning parameter.

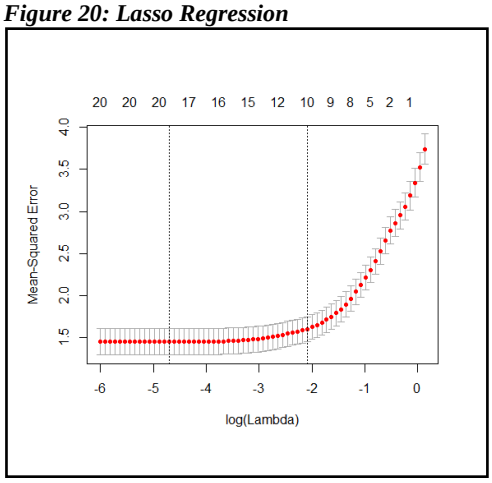
Figure 19: Ridge Regression



- Ridge Regression keeps all variables and shrinks the coefficients towards zero. We pick the best value for lambda through cross-validation.

Ridge regression includes all predictor variables in the final model. The Lasso overcomes this limitation and manages to reduce the predictors.

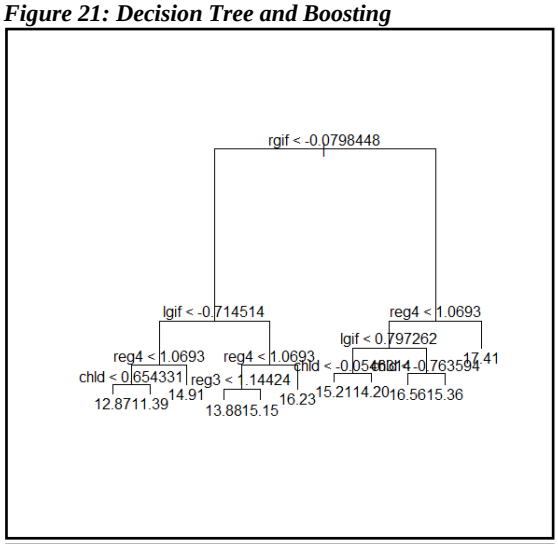
Figure 20: Lasso Regression

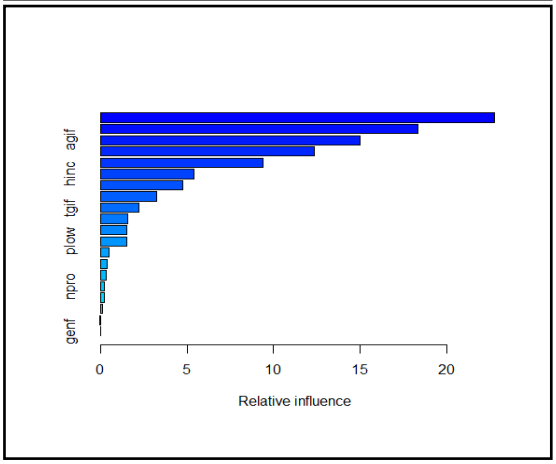


- The output on the Lasso gives an indication of how many variables we need to add to our model to achieve the lowest mean squared error. This model appears to hit a good spot at around 15 predictors.

We had a lot of success with the various elements of Decision Tree Models in our classification analysis. The models we have worked through thus far have produced decent results but have fallen short.

Figure 21: Decision Tree and Boosting





- The decision Tree on the left gives us a good feeling for what variables we need to fit for our predictive model. We know that decision trees overfit data, but we also developed bagging and boosting models to reduce the bias and variance. The Relative influence plot is directly below the decision tree.

The criteria for selecting the best prediction model for donation amount (DAMT) will be Mean Squared Prediction Error and Standard Error. The Mean Squared Prediction Error measures the expected squared distance between what your predictor predicts for a specific value and what the real value actually is. It serves as a gauge of quality for Predictive Models. The Standard Error of the regression is a goodness of fit metric indicating how well your model fits the data. The standard error of the regression provides the measure of the distance for our data points from the regression line. Our goal is to have the best possible predictive model for DAMT so we will base the final selection decision on the Mean Squared Prediction Error.

Prediction model Selection:

The table below has all the prediction models we incorporated into the analysis organized by their Mean Square Prediction Errors with the best performing model listed first and descending from there.

Figure 10: Prediction Model for DAMT Summary

Model	Model Mean Prediction Error	Model Standard error
Boosting	1.540	0.167
OLS Regression	1.612	0.166
Lasso Regression	1.614	0.165
Forward Selection	1.615	0.165
Ridge Regression	1.636	0.167
OLS Best Variable Selection	1.640	0.164
Random Forest	1.656	0.172
Bagging	1.726	0.176
Decision Tree	2.241	0.192

The boosting model is once again the best performer with respect to predicting potential donation amounts of our donors. We will proceed with this model for our final prediction of donation amount (DAMT). While the Standard Error wasn't the best on this model, it was within the range of the other models we used in our analysis. The one exception would be the Decision Tree, but that isn't entirely surprising, given that Decision Tree's are typically better suited for classification and exploratory data analysis. However, it's interesting to note that the incorporation of Boosting, which is a predictive approach for improving the results of Decision Tree's, produced the best working model according to our metrics.

Final Profit:

The direct marketing campaign whether successful or not will cost the charity money. We now have two distinct models we can use to develop our machine learning model to improve the cost-effectiveness of the charity's direct marketing campaign. Based on the model, we will mail to the 321 highest posterior probable donors. This results in an estimated final profit of $4,657.43 dollars and an average donation amount of $14.51.

Conclusion:
This assignment was designed to develop a machine learning model to improve the cost-effectiveness of a charity's direct marketing campaign. We fit nine different classification models in to find the best performer in terms of future donors, and nine different regression models which helped us predict the donation amount to be expected. The classification model which incorporated boosting was the best performer for predicting donors. Similarly, the prediction model which used gradient boosting was also the best performer when it came to predicting donation amounts. The results of this machine learning model should help the charity better target donors and save money on future mailings

Future work and learning:
While OLS and Logistic regression are time-tested as viable tools, it's clear they aren't as effective at modeling data outside what one would consider linear. Perhaps we could spend more time on the exploratory phase with similar models in an effort to assess their non-linearity and move towards more robust modeling techniques from the start of the analysis, rather than starting with modeling techniques we preemptively deem to be inefficient for the data at hand.