Final Project

MSDS 422 - Section 56
September 1, 2018

1 Introduction

In this report I develop a machine learning model to improve the cost-effectiveness of a charitable organization's direct marketing campaigns to previous donors. I approach this task in two stages. First, I explore various classification models as a means of identifying likely donors to whom the organization should direct their marketing campaigns to maximize profits. Second, I explore various prediction models to predict how much a donor will actually give. Each of the machine learning models tested are described in detail. All models are compared analytically (i.e., classification compared with classification, prediction compared with prediction) to select the best overall model of each type. Finally, the two champion models are applied to a test data set.

2 Data Exploration

The charity data set used in this analysis is provided by the MSDS 422 class taught by Professor Bastian and was obtained from Canvas. It contains 8009 observations of 23 total variables of which 20 can be used as predictor variables:

- ID number (an identification number, not used as a predictor)
- region (reg1, reg2, reg3, reg4--observations with 0 in each of these belong to region 5)
- home owner (home)
- number of children (chld)
- category of household income (hinc)
- gender (genf)
- health rating (wrat)
- average home value (avhv)
- median and average family income (incm and inca)
- percent categorized as low income (plow)
- number of lifetime promotions (npro)
- lifetime gifts to date in dollars (tgif)
- largest gift to date in dollars (lgif)
- most recent gift in dollars (rgif)
- average gift in dollars (agif)
- months since last donation (tdon)
- months between first and second donation (tlag)

The classification response variable donr indicates whether the observation is a donor or not (binary variable of one if a donor and zero if not a donor). The prediction response variable damt is the donation amount (for donors only) in dollars.

Of the 8009 observations in the data set, 3984 observations are in the training set, 2018 observations are in the validation set, and 2007 observations are in the test set. To account for the differences in scale and variance of the different predictor variables, each data set was standardized to have a mean of zero and standard deviation of one. No value imputation was needed since there are no missing values for predictor variables.

2.1 Exploratory Data Analysis

For my exploratory data analysis (EDA), I examined each of the variables in the data set in detail. First, I visualized the data structure and looked at the distributions of each of the variables. Figures 1 and 2 in the appendix provide an overview of the histograms I created.

As shown in these histograms, several of the variables are right skewed (e.g., incm, inca, avhv, rgif, tgif, tlag, agif) and could benefit from log transformations. I explored transforming these variables to ensure more normal distributions and examine the impact of those transformations on my models later in this analysis. While several of

the variables have outliers, I have chosen to conduct log transformations instead of trimming. Additionally, I used several tree-based model approaches that are robust to the presence of outliers (e.g., gradient boosting machines, random forest, decision trees).

Next, I looked at correlations between variables to determine what predictor variables may have the greatest impact on the target variable (donr) and which predictors may be correlated with one another. The correlation plot in Figure 3 shows the correlations for the donr target variable and the predictors. The correlation plot in Figure 4 shows the correlations for the damt variable when donr=1.

Figure 3. Charity data set correlations with donr

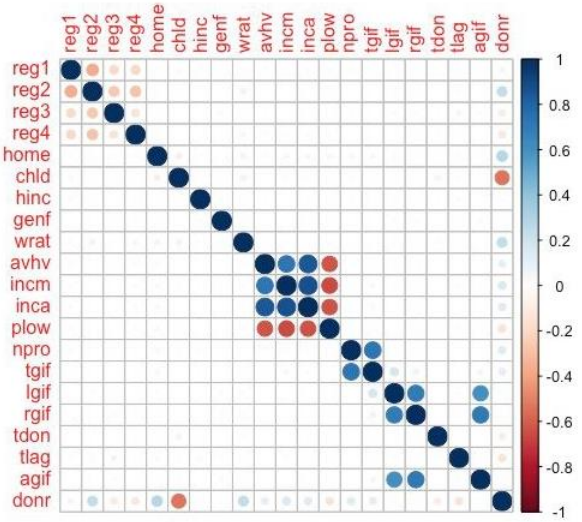
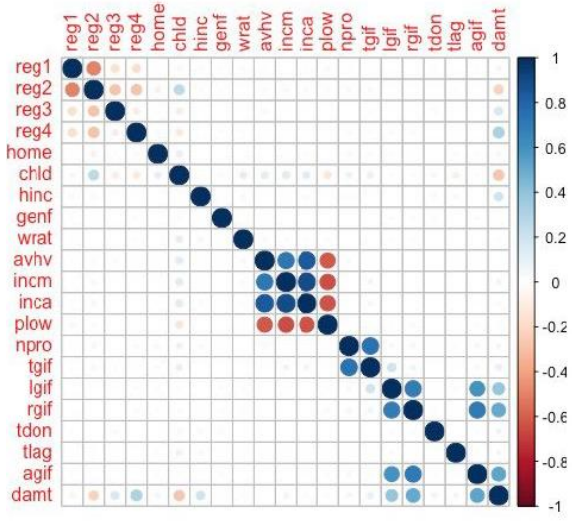


Figure 4. Charity data set correlations with damt (donr=1)



In the correlation plot for predictor variables and the target variable (donr) in Figure 3, I can see that chld is the variable with the largest correlation (negative) to donr. Other variables such as reg2, home and wrat appear to have slight positive correlations to donr. Interestingly, no variables appeared to have strong positive correlations to donr. Several predictor variables have strong correlations to other predictor variables such as incm, inca, avhv and plow. This will need to be considered when building certain models as multicollinearity may affect the interpretability of my coefficients.

In the correlation plot for damt for donors (Figure 4), I can see that the variables with the greatest correlation with the amount donated (damt) are those related to any prior gifts received (e.g., rgif, agif, lgif) and reg4. Interestingly, chld and reg2 are moderately negatively correlated with the amount donated. As with the donr correlations, I see that there are predictor variables for damt that are correlated with one another (npro and tgif, lgif and rgif) and I will need to look out for multicollinearity in my model coefficients.

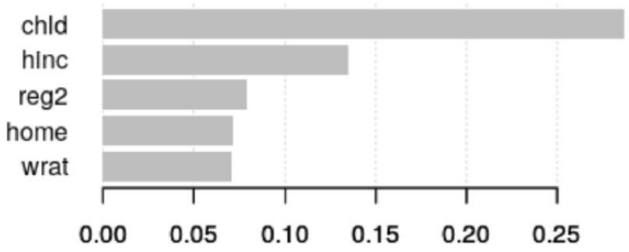
I also conducted bivariate analysis to look at the differences between donors and non-donors, and saw several differences between groups. Specifically, donors appear to be more likely to reside in reg 1 and 2, own a home (home), have a higher wealth rating (wrat), live in a neighborhood with a higher average home value (avhv), live in a neighborhood with higher median family income and average family income (incm, inca), have received a higher number of promotions in their lifetime (npro) and have a higher amount of lifetime gifts to date (tgif). Donors are less likely to reside in reg 3 and 4 and to live in a neighborhood with low income households (plow), as shown below:

Figure 5. Bivariate analysis: mean values by donor category

	Mean Variable Values by Donor Category	
	Non-Donor (donr=0)	Donor (donr=1)
reg1	0.182	0.228
reg2	0.219	0.453
reg3	0.158	0.089
reg4	0.178	0.092
home	0.790	0.976
chld	2.330	0.830
hinc	3.907	3.985
genf	0.613	0.596
wrat	6.474	7.631
avhv	176.281	194.061
incm	40.304	48.261
inca	53.608	60.652
plow	15.611	11.860
npro	57.506	65.740
tgif	106.829	126.631
lgif	22.389	23.994
rgif	15.367	15.726
tdon	19.353	18.278
tlag	6.810	5.795
agif	11.598	11.720

Additionally, I used XGBoost (a gradient boosting approach) to provide output regarding variable importance. Below is an overview of the top five variables that XGBoost chosen as most important. These findings are interesting in that they highlight two variables that did not appear to be correlated with donr (hinc and wrat) in my plots but still appear to be important:

Figure 6. Variable importance per XGBoost



3 Analysis

In this section, I describe in detail all the models fit and tested, both for classification and prediction. I assess their capabilities based on goodness-of-fit metrics in the training set and performance metrics in the validation set. Finally, I select one classification model and one prediction model to apply to the test set.

3.1 Classification Models for DONR

I tested eleven different classification models including:

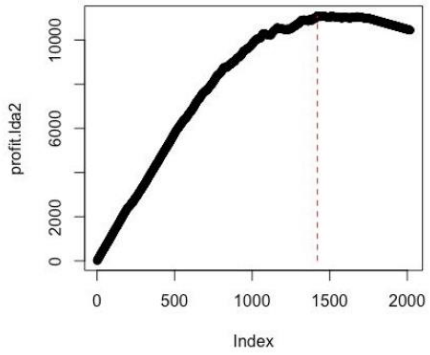
- Linear Discriminant Analysis (LDA with variable selection)
- Quadratic Discriminant Analysis (QDA)
- Generalized Additive Model (GAM)
- Multiple Logistic Regression (MLR) (with variable selection)
- Gradient Boosting Machines (two models, with and without log transformations)
- XGBoost (a well-known gradient boosting approach using parallel tree boosting)
- Random Forest (three models)
- Neural Net

The following sections describe these models in greater detail.

3.1.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a classification method that can be more stable than logistic regression when classes are well-separated, and n is small. It does assume that the distributions of the predictor variables are Gaussian so I have used scaled data with a mean equal to zero in this model. My approach created an LDA model using only the top five predictors (chld, hinc, home, reg2, wrat) that Ire identified as important in my XGBoost variable importance output. The red-dashed vertical line represents the optimal number of mailings for maximum profit based on the model.

Figure 7. Expected profits by number of mailings (LDA)



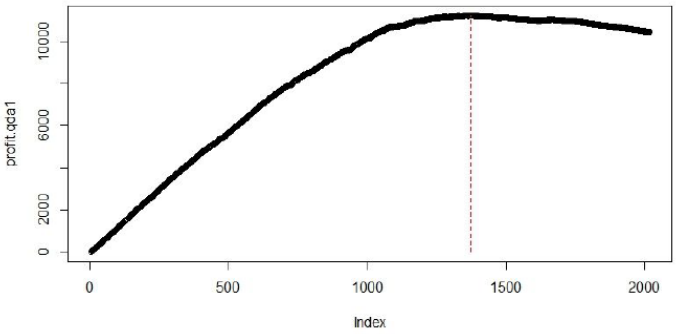
Overall, this LDA approach did not appear to outperform the LDA model provided by the professor. By applying the LDA model to the validation set, I can see that the accuracy rate is 75.37%. The optimal number of mailings from this model is 1420, resulting in an expected profit of $11,094.50. This was my lowest performing model.

3.1.2 Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) provides an alternative linear discriminant analysis approach. Like LDA, the QDA classifier results from assuming the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix.

After applying a QDA model to the validation set, the classification accuracy rate is 77.95%. The highest number of mailings from this model in Figure 8 is 1372, which gives a profit of $11,219.50.

Figure 8. Expected profits by number of mailings (QDA)



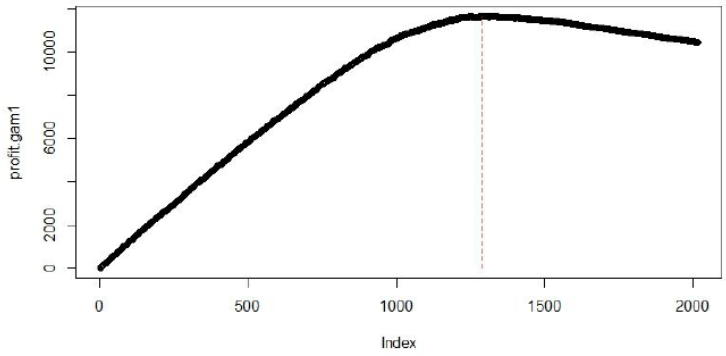
3.1.3 General Additive Model

Generalized additive models (GAMs) provide a general framework for extending a standard linear model by allowing

non-linear functions of each of the variables, while maintaining additivity. Just like linear models, GAM's additivity can be applied to both quantitative and qualitative responses. GAMs allow us to model non-linear relationships that standard linear regression will miss. The non-linear fits can potentially make more accurate predictions for the response variable. With GAMs, I do not need to manually try out many different transformations on each variable individually.

The classification accuracy rate is 83.8% for this model. The highest number of mailings in Figure 9 is 1290, which gives a profit of $11,644.50. This is an improvement over both my LDA and QDA models.

Figure 9. Expected profits by number of mailings (GAM)



3.1.4 Multiple Logistic Regression

A logistic model is a common statistical method used to predict a binary dependent variable. The result of a logistic regression model is the probability of a positive response, where "positive" is defined by the user (e.g., "yes" response). While my professor provided a logistic regression model that contained all the predictor variables, I undertook a logistic regression approach with variable selection in an attempt to improve upon this baseline model.

Multicollinearity is an issue that occurs when predictor variables are correlated with one another. This results in coefficient estimates that are unstable and difficult to interpret. While not detrimental to the performance of a model, these issues with the coefficients make it an undesirable trait. Multicollinearity can be checked by calculating the variance inflation factor (VIF) of each coefficient, and the largest five VIF scores for the logistic model are shown in Table 1 below:

Table 1. Logistic model VIF scores

Coefficient	VIF Score
inca	6.563
incm	4.348
avhv	3.530
rgif	2.631
lgif	2.165

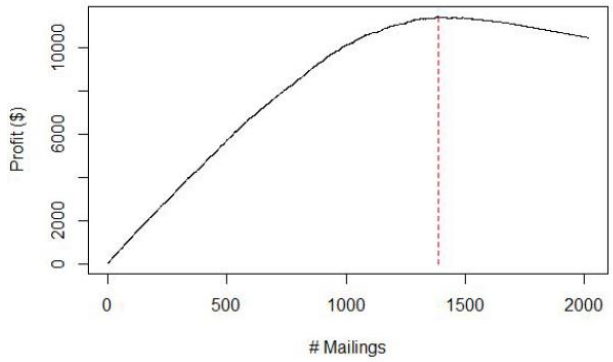
I consider a VIF above five to be cause for concern. Because inca has a VIF higher than five, I investigated further. In the logistic model, inca is not statistically significant in the model. Therefore, I removed inca from the model and fit the multiple logistic regression model again. The coefficients and their estimates for the logistic regression model is given in Table 2:

Table 2. Logistic model coefficient estimates

reg1	reg2	reg3	reg4	home	chld	hinc	genf	wrat	avhv	incm	plow	npro	tgif	lgif	rgif	tdon	tlag	agif
56	1.21	0.02	-0.01	1.12	-1.96	0.08	-0.03	0.82	0.05	0.37	-0.19	0.38	0.14	-0.03	0.02	-0.23	-0.46	0.03

Applying this model to the validation set resulted in a classification accuracy of 78.7%. Based on the probabilities returned, I computed the expected profits as more mailings are sent out, which can be seen in Figure 10 below.

Figure 10. Expected profits by number of mailings (Logistic Regression Model)



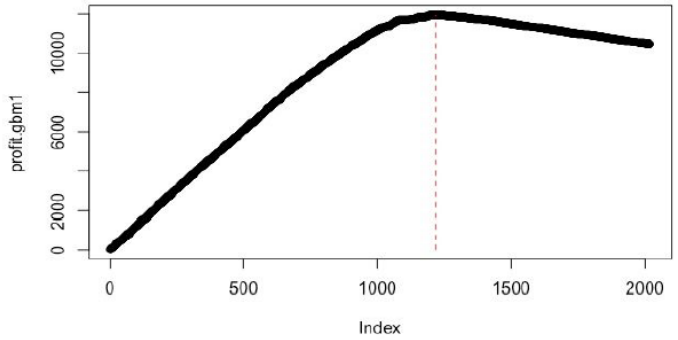
The optimal number of mailings from this model is 1385, resulting in an expected profit of $11,425.50. Unfortunately, this was not an improvement over the baseline logistic regression model provided in the project code with an optimal number of mailings at 1291 and an expected profit of $11,642.50.

3.1.5 Gradient Boosting Machine 1

Next, I decided to undertake a gradient boosting approach. Boosting creates one prediction model from an aggregation of decision trees. These trees are grown sequentially using the information from the previously grown trees. Each tree utilizes a subset of the original data set as opposed to bootstrapping.

The first gradient boosting machine I created utilized all the predictor variables and one transformation of the hinc predictor. After applying the gradient boosting model to the validation set, the classification accuracy rate is 88.4%. The highest number of mailings from this model in Figure 11 is 1219, which gives a profit of $11,946. This was the best performing model out of all the models created.

Figure 11. Expected profits by number of mailings (Gradient Boosting Machine 1)

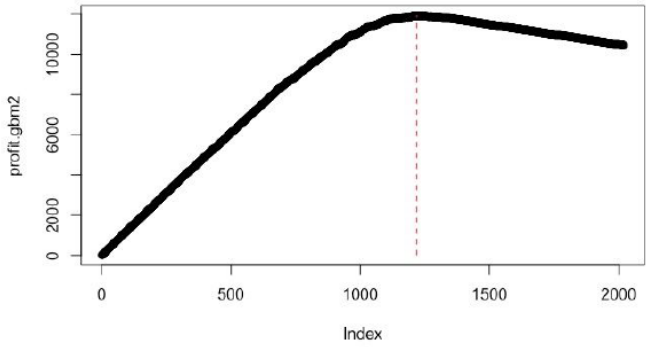


3.1.6 Gradient Boosting Machine 2

I also attempted to improve my original gradient boosting machine by removing scaling and log transforming all the right skewed predictor variables (hinc, incm, inca, plow, lgif, rgif, avhv). This resulted in a model that performed

similarly to my original gradient boosting machine with an classification accuracy rate of 88.35%. The highest number of mailings from this model in Figure 12 is 1218, which gives a profit of $11,933.50.

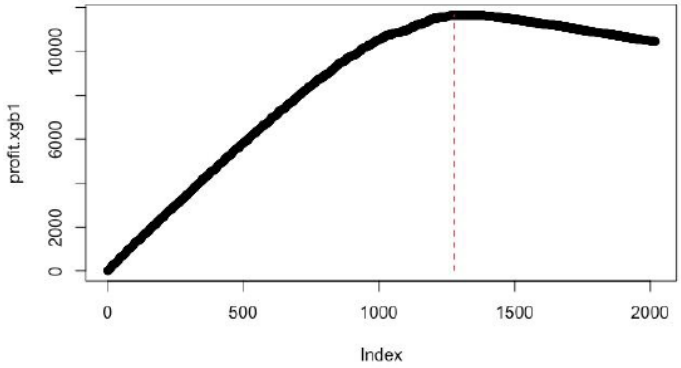
Figure 12. Expected profits by number of mailings (Gradient Boosting Machine 2)



3.1.7 XGBoost

Next, I utilized a variant of a gradient boosting approach called XGBoost, which stands for eXtreme Gradient Boosting. While this approach often performs better than GBM, this model did not outperform the best-performing GBM1 model. The highest number of mailings from this model in Figure 13 is 1236, which results in a profit of $11,870.50. The accuracy rate of this model is 86.3%.

Figure 13. Expected profits by number of mailings (XGBoost)

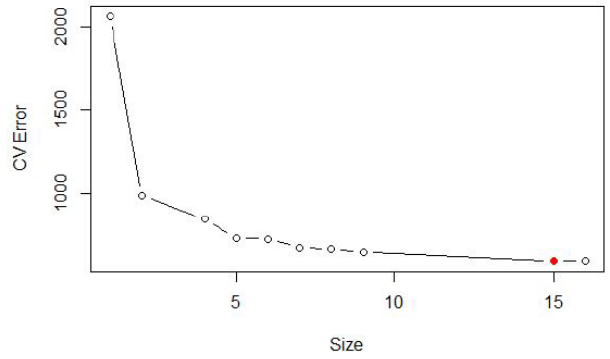


3.1.8 Decision Trees & Random Forest

Decision trees for classification problems predict each observation to belong to the most commonly occurring class of training observations in the region to which it belongs. Classification error rate is used to determine branch splits, supplemented by measures of node purity (Gini index, entropy). Because they closely mirror human decision-making and can be easily represented graphically, decision trees are very easy to explain and interpret.

Next, I fit a classification tree to the training data and used cross-validation to determine the optimal level of tree complexity. Figure 14 shows the trend in cross-validation error with changes in tree complexity. The red dot represents the optimal tree size.

Figure 14. Classification tree cross-validation error versus tree size



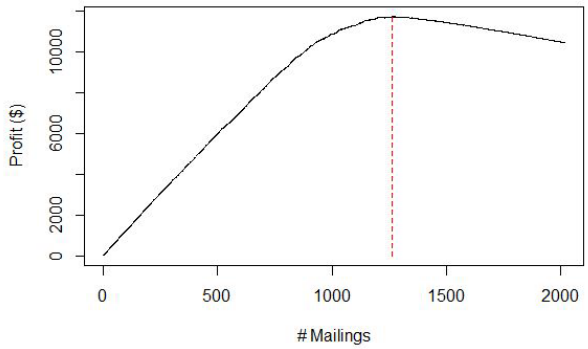
The classification tree with 15 terminal nodes resulted in a classification accuracy of 85.28%. Classification trees can often see increased accuracy with the use of random forests. With random forests, I build several decision trees with bootstrapped training samples. Each time a split is considered, a random sample of predictors is chosen as split candidates, and only one is ultimately selected. This helps prevent the strongest predictor from always being selected, thereby decorrelating the bagged trees, reducing the variance and increasing the reliability. I tested three different random forest models, fitting 500 trees each. The sample sizes for split candidate predictors Ire 20 (i.e., same as bagging), 12, and the default of the square root of the number of predictors which amounted to 4. The classification accuracies are shown in Table 2.

Table 2. Random forest model accuracies

Candidate Predictor Sample Size	Classification Accuracy
20	90.09%
12	88.69%
4	90.09%

The best decision tree model is the random forest using a sample size of 12 predictors, which improved over the original classification tree. Applying this model to the validation set resulted in a classification accuracy of 88.69%. Based on the probabilities returned, I computed the expected profits as more mailings are sent out, which can be seen in Figure 15. The red-dashed vertical line represents the optimal number of mailings for maximum profit based on the logistic regression model.

Figure 15. Expected profits by number of mailings (Random Forest Model)

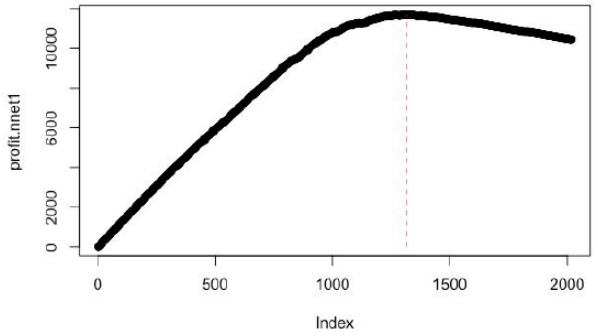


The optimal number of mailings from this model is 1211, resulting in an expected profit of $11,701.

3.1.10 Neural Net

As my final approach, I deployed a classification model based on neural networks. Neural networks utilize algorithms that are loosely modeled after the workings of the human brain. One can think of these as an extension of linear or logistic regression in that there are hidden layers inserted between the inputs and outputs with non-linear activation units. This approach did not outperform my best model (GBM1). The accuracy of the model is 83.15%. The optimal number of mailings from this model is 1317, resulting in an expected profit of $11,692.00.

Figure 16. Expected profits by number of mailings (Neural Net)



3.1.11 Comparison of Classification Models

Performance metrics from the validation set are given for each model in Table 3. Accuracy refers the overall proportion of correct classifications (donor and non-donor). Sensitivity is the true positive rate and measures the ability to correctly identify donors. Specificity is the true negative rate and measures the ability to correctly identify non-donors. Precision is the positive predictive value and measures the percentage of correctly identified donors out of all predicted donors--in other words, it's a measure of my confidence in the model's ability to correctly predict a donor and not a false positive.

Table 3. Classification model performance on validation set comparison

Model	Accuracy	Sensitivity	Specificity	Precision
LDA	0.754	0.962	0.550	0.677
QDA	0.780	0.964	0.599	0.702
GAM	0.838	0.981	0.697	0.760
MLR	0.787	0.980	0.602	0.705
GBM1	0.884	0.993	0.777	0.814
GBM2	0.884	0.992	0.777	0.814
XGB	0.863	0.994	0.734	0.786
DT	0.852	0.943	0.764	0.797
RF*	0.873	0.972	0.776	0.810
NN	0.832	0.989	0.677	0.750

*best performing random forest model

The original gradient boosting machine (GBM1) is the winner in almost every category. I also performed operational validation by calculating what each model would consider the optimal number of mailings to send and the associated expected profits. These results are shown in Table 4.

Table 4. Optimal number of mailings and expected profit by model

Model	Optimal # Mailings	Expected Profit ($)
LDA	1,420	11,094.50
QDA	1,372	11,219.50
GAM	1,291	11,642.50
MLR	1,389	11,417.50
GBM1	1,219	11,946.00
GBM2	1,218	11,933.00
XGB	1,236	11,870.50
DT	1,487	11,381.00
RF*	1,211	11,701.00
NN	1,317	11,692.00

*best performing random forest model

Again, gradient boosting machine (GBM1) is the winner by returning the highest expected profit from the least amount of mailings. I selected gradient boosting as my champion classification model.

3.2 Prediction Models for DAMT

I tested six different prediction methods including best subset selection (BS), ridge regression (Ridge), lasso regression (Lasso), principal components regression (PCR), random forest (RF) and gradient boosting machine (GBM). The following sections describe these models in greater detail.

3.2.1 Best Subset Selection

Linear regression simply models the relationship between a response and one or more explanatory variables in a linear fashion. Like logistic regression, the parameters are estimated using maximum likelihood estimation to choose values that are most likely given the data. Best subset selection is a form of variable selection to help determine which variables to include in the model. As was mentioned above during my EDA, skewed predictor variables exist in the charity data set, which can cause violations to the assumptions of linear regression. Several types of transformations including log, Box-Cox and power were considered and applied to help normalize skewed predictor variables, such as avhv, incm, inca, tgif, lgif, rgif, and agif. Additionally, two new predictor variables Ire created in an attempt to try and gain more predictive power. The first was an indicator of whether the observation had zero kids or not (chld0). The second was an indicator of whether the observation had a health rating greater than five (wratHIGH).

The best models Ire found using stepwise, forward and backward variable selection methods. The best model based on Bayes' Information Criterion (BIC) contained 13 predictor variables for all three methods. The best model based on Mallow's Cp (Cp) contained 16 predictor variables for all three methods. The best model based on adjusted R^2 contained 17 predictor variables for forward and backward selection, and 18 predictor variables for stepwise selection.

The best simple linear regression model (i.e., regression model with only one predictor) always used the predictor variable lgif. I compared this linear model to those with 13, 16, 17 and 18 predictors. Each variable selection

method resulted in the exact same coefficients except for the stepwise method for 17 predictors. Table 5 displays the validation error for each model.

Table 5. Best subset model validation errors

# Coefficients	Mean Validation Error	SE Validation Error
1	2.6808	0.2131
13	1.5180	0.1585
16	1.5160	0.1587
17 (forward/backwards)	1.5157	0.1590
17 (stepwise)	1.5738	0.1640
18	1.5159	0.1590

From this I see that best subset selection recommends the linear model with 17 predictor variables chosen by either forward or backward selection. The coefficients and their estimates for the best subset regression model are below:

Table 6. Best subset model coefficient estimates

reg1	reg2	reg3	reg4	home	chld	hinc	genf	wrat	incm	plow	tgif	lgif	rgif	tdon	agif	wrat HIGH
0.04	-0.08	0.31	0.64	0.25	-0.62	0.51	-0.06	-0.40	0.45	0.40	0.21	0.47	0.44	0.06	0.34	0.47

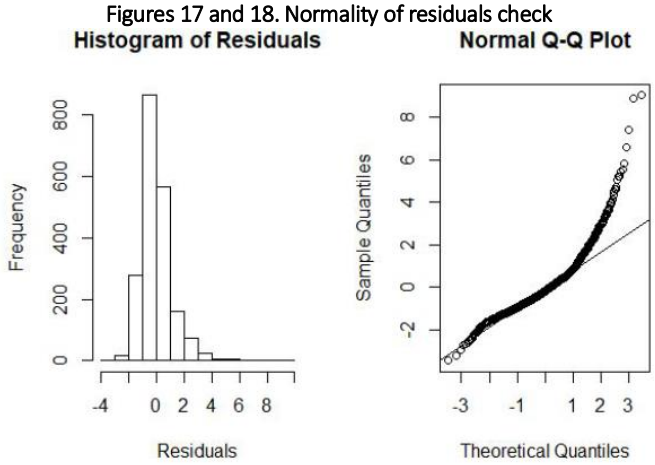
Although this algorithm determined the optimal number of predictors in terms of criterion such as adjusted R_2 or validation error, the model is not without its drawbacks. As with logistic regression, I calculated the VIF scores of this model, and the top five are shown in Table 7:

Table 7. Best subset model VIF scores

Coefficient	VIF Score
lgif	5.788
rgif	3.979
agif	3.667
wratHIGH	3.158
wrat	3.127

Variable lgif has a VIF score over five, which I consider problematic. However, given that I used the best subset algorithm to tell us which variables to include in the model, I'll allow this while acknowledging the limitations it imposes. Figures 17 and 18 display residual diagnostic plots of this model.

Figures 17 and 18. Normality of residuals check



The residuals don't appear to be fully normally distributed, which will be kept in mind as I compare these results to other models.

3.2.2 Ridge Regression

Ridge regression is a shrinkage method of variable selection. I fit a linear regression model containing all the predictors and use ridge regression to constrain or shrink the coefficient values of some (or all) of the predictors closer to zero. Ridge regression will never set coefficient estimates to zero, so the resulting model is still the full model. The amount to which coefficients are shrunk towards zero is dependent upon the tuning parameter lambda. Large lambda values correspond to lots of shrinkage. Small lambda values correspond to little shrinkage and estimates like those of least squares regression.

I fit a ridge regression model using a grid of lambda values (i.e., fit several models each with a different tuning parameter value). Using cross-validation, I obtained values for lambda to be used in two different ridge regression models. The first used the value of lambda (0.1108) that resulted in the smallest cross-validation error. The second used the largest value of lambda (0.7120) such that the cross-validation error was within one standard error of the minimum. Table 8 displays the validation error associated with each model.

Table 8. Ridge regression model validation errors

Lambda	Mean Validation Error	SE Validation Error
0.1108 (min)	1.873	0.171
0.7120 (1se)	1.995	0.179

The ridge regression model with the lambda which resulted in the smallest cross-validation error (0.1108) resulted in the best accuracy on the validation set.

3.2.3 Lasso Regression

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model.

To build this model, a coefficient plot was created to see whether any of the coefficients will be equal to zero.

Then, cross-validation was performed utilizing both lambda being at minimum, and lambda being one standard error from the minimum; last, the associated test error was computed.

Table 9. Lasso regression model validation errors

Lambda	Mean Validation Error	SE Validation Error
0.0026 (min)	1.861	0.169
0.0898 (1se)	1.971	0.171

Of the 18 variables available, only one (wrat) was zero. This means the lasso model with lambda chosen by cross-validation contains only 17 variables. The lasso regression model with the lambda which resulted in the smallest cross-validation error resulted in the best accuracy on the validation set.

3.2.4 Principal Components Regression (PCR)

Principal components regression (PCR) is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large, so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors.

In building the PCR model, the following parameters Ire set: scale = FALSE (TRUE has the effect of standardizing each predictor but these predictors have already been scaled) and validation="CV" (causes pcr() to compute the ten-fold cross-validation error for each possible value of M, the number of principal components used). Using the validationplot(), the cross-validation scores are shown in Figure 19 below:

Figure 19. Cross-validation scores

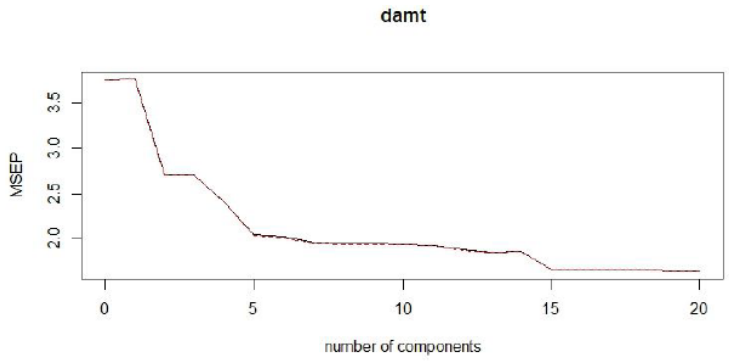


Table 10. PCR validation errors

Mean Validation Error	SE Validation Error
1.876	0.170

3.2.5 Gradient Boosting Machine

For the purposes of predicting damt, I also utilized a gradient boosting machine approach. This approach resulted in the lowest mean prediction error (1.379) of any of the prediction models I created. The gradient boosting model returned variable importance results that showed that variables related to past donations Ire the most important in predicting damt, as shown below:

Table 11. GBM Variable Importance

var	relative importance
rgif	20.4694348
lgif	16.5282028
agif	14.1142232
reg4	11.0638585
chld	8.9535594
wrat	5.6330985
hinc	5.5466081
reg3	3.1222256
tgif	3.1153448
incm	2.4499205
plow	2.2791215
reg2	1.4419096
avhv	1.1896465
inca	1.0304069
npro	0.844876
home	0.6859582
tdon	0.6551597
tlag	0.3058074
reg1	0.2724395
genf	0.1798603

The GBM model returned the following errors:

Table 11. GBM validation errors

Mean Validation Error	SE Validation Error
1.379	0.161

3.2.6 Random Forest

Additionally, I created a random forest model to predict damt. This model utilized all the predictor variables including income squared ($hinc^2$). This model performed third best among the models I created and received the following errors:

Table 12. Random forest validation errors

Mean Validation Error	SE Validation Error
1.664	0.174

3.2.7 Comparison of Prediction Models

Table 13. Prediction model performance on validation set comparison

Model	Mean Validation Error	SE Validation Error
BS	1.516	0.159
Ridge*	1.873	0.171
Lasso*	1.861	0.169
PCR	1.876	0.170
GBM	1.379	0.161
RF	1.664	0.174

*best performing model among those tested for that approach

As with the classification model, gradient boosting machine is the winner by returning the lowest mean and SE validation errors. I selected gradient boosting as my champion model for predicting damt.

4 Results

Of the classification models I fit to the training data and validated, my original gradient boosting model (GBM1) was the winner. It had classification accuracy rate of 88.4%. The highest number of mailings from this model was 1219, which resulted in a profit of $11,946.

The top performing prediction model based on the validation set was also my gradient boosting model, with a mean validation error of 1.379 and a SE validation error of 0.161.

5 Conclusion

In this analysis, I developed and deployed several different machine learning approaches to improve the cost-effectiveness of a charitable organization's direct marketing campaigns to previous donors. Several classification models Ire tested and compared to identify likely donors, with gradient boosting resulting in the best model. I also created and compared a variety of machine learning approaches to predict the amount a donor would give, and gradient boosting again performed best out of all models tested.

6 Appendix

Figure 1. Charity data set variable histograms

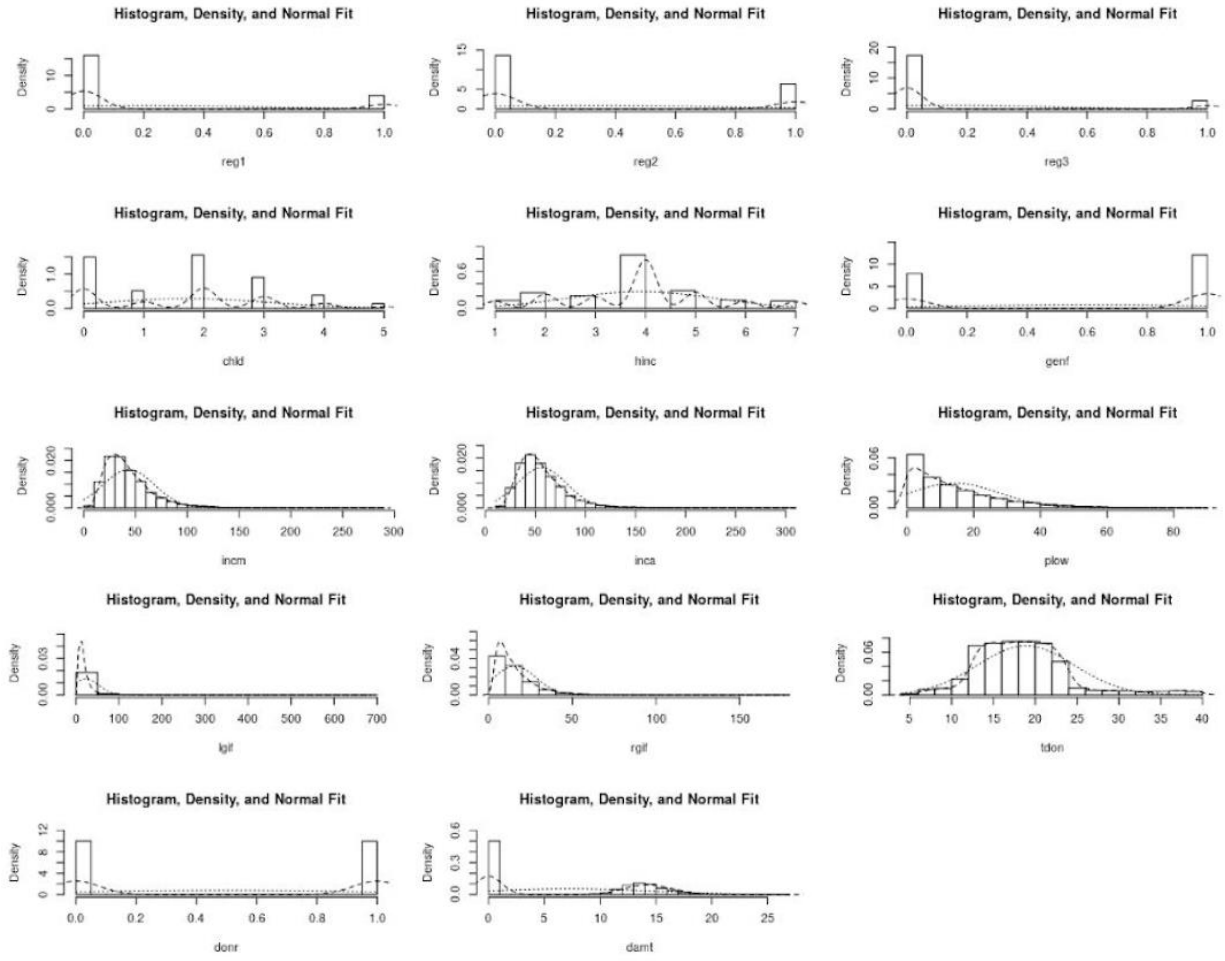
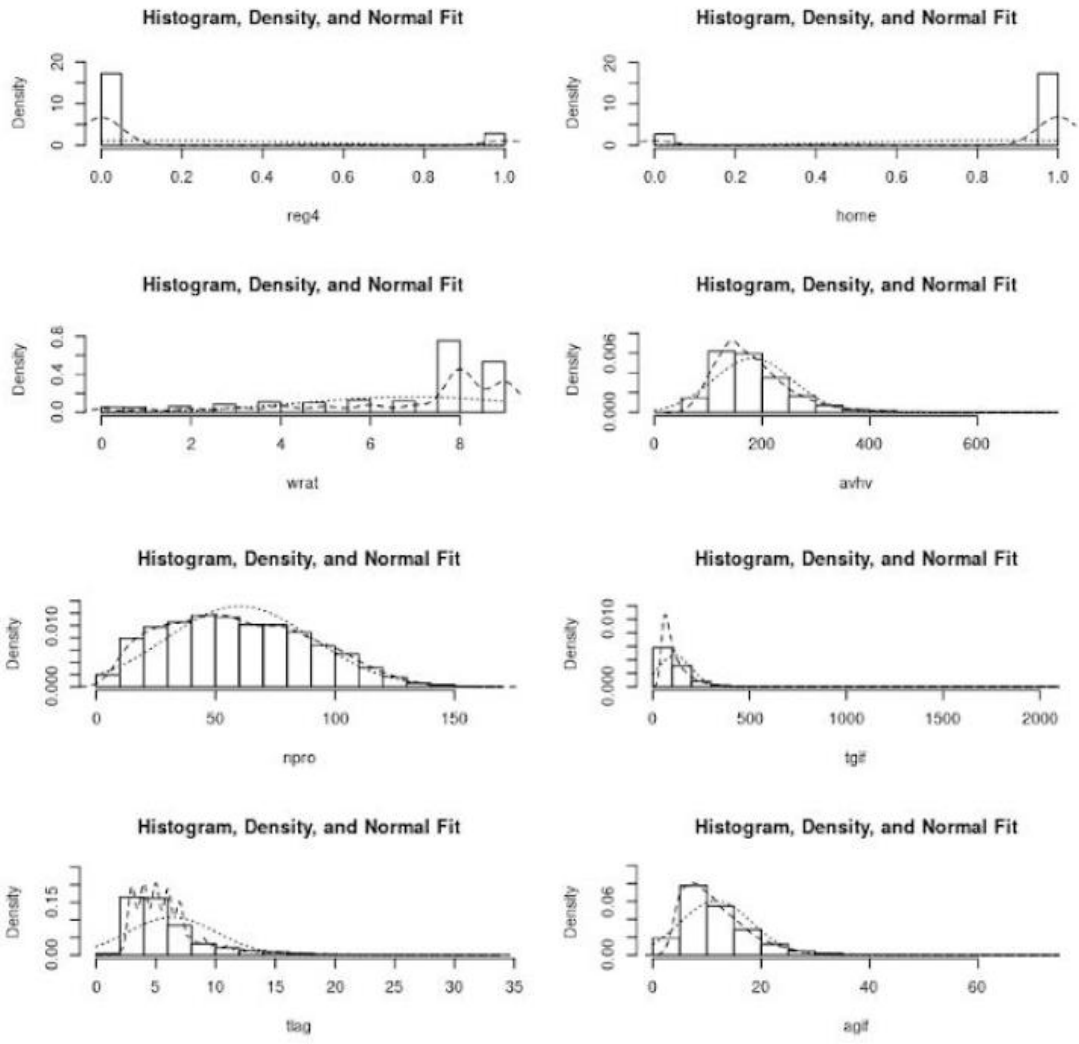


Figure 2. Charity data set additional variable histograms



7 References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York: Springer.

U. (2016, August 10). A Quick Introduction to Neural Networks. Retrieved from https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/