Course Project: Charity

PREDICT 422 Section 58

Scott M. Morgan

INTRODUCTION

The purpose of this report is to build classification and prediction models so a charitable organization can improve the cost-effectiveness of their direct marketing campaigns to previous donors. A classification model will first be developed using data from the organization's most recent campaign that effectively captures likely donors so that the expected net profit is maximized. Following this, a prediction model to forecast the expected gift amounts from donors will be constructed. The data for this portion will consist of the records for donors only. To accomplish both these objectives, RStudio is used to perform the analysis and construct models. Several parametric and non-parametric techniques will be utilized as well as manual and automated variable selection methods.

RESULTS

In the subsequent sections, a series of classification and prediction models are generated and evaluated. The functionality within RStudio is used to perform a brief exploratory data analysis (EDA) to build an understanding of potential predictor variables and their relationship to the responses. Following this, variables are examined for deficiencies such as missing data and outliers as a precursor to preparing the data set for modeling through imputation, elimination and/or transformation. Finally, the models are generated and evaluated.

EXPLORATORY DATA ANALYSIS (EDA). An assortment of tools are deployed to perform the initial EDA before cleaning the data and ultimately constructing the classification and predictive models. I begin the analysis by examining the variables in the entire data set using the *describe.by* function. The summary for the entire raw data set is provided in Table 1 below.

Table 1: Charity Raw Data Set Summary

Variable	Count	Mean	Std. Dev.	Median	Min	Max	Range	Skew	Kurtosis
REG1	8009	0.2	0.4	0	0	1	1	1.5	0.24
REG2	8009	0.32	0.47	0	0	1	1	0.78	-1.4
REG3	8009	0.13	0.34	0	0	1	1	2.15	2.63
REG4	8009	0.14	0.35	0	0	1	1	2.08	2.33
HOME	8009	0.87	0.34	1	0	1	1	-2.16	2.64
CHLD	8009	1.72	1.4	2	0	5	5	0.27	-0.8
HINC	8009	3.91	1.47	4	1	7	6	0.01	-0.09
GENF	8009	0.61	0.49	1	0	1	1	-0.43	-1.81
WRAT	8009	6.91	2.43	8	0	9	9	-1.35	0.79
AVHV	8009	182.65	72.72	169	48	710	662	1.54	4.49
INCM	8009	43.47	24.71	38	3	287	284	2.05	8.31
INCA	8009	56.43	24.82	51	12	305	293	1.94	7.87
PLOW	8009	14.23	13.41	10	0	87	87	1.36	1.89
NPRO	8009	60.03	30.35	58	2	164	162	0.31	-0.62
TGIF	8009	113.07	85.48	89	23	2057	2034	6.55	107.52
LGIF	8009	22.94	29.95	16	3	681	678	7.81	110.38
RGIF	8009	15.66	12.43	12	1	173	172	2.63	13.92
TDON	8009	18.86	5.78	18	5	40	35	1.1	2.12
TLAG	8009	6.36	3.7	5	1	34	33	2.42	8.41
AGIF	8009	11.68	6.57	10.23	1.29	72.27	70.98	1.78	6.02
DONR	6002	0.5	0.5	0	0	1	1	0	-2
DAMT	6002	7.21	7.36	0	0	27	27	0.12	-1.83

The entire data set contains 8009 observations and 24 variables; 1 variable is a unique identifier and the other separates the data into train, validation and test sets. Both are excluded from the above table and proceeding analysis. When partitioned by PART (not shown), the data set consists of 3984 training observations, 2018 validation observations, and 2007 test observations. Weighted sampling was used, over-representing the responders so that the training and validation samples have approximately

equal numbers of donors and non-donors. The response rate in the test sample has the more typical 10% response rate and thus the number of mailings will be scaled appropriately.

There are a total of 11 continuous variables, 3 count variables and 6 categorical independent variables. Each record represents a male or female who either has or has not donated to the charity and how much was donated. Each record has 2 target variables that will be predicted. The first target variable, DONR, is a “1” or “0”. A “1” indicates that the person has donated in the past. A “0” indicates that the person has not donated. The second target variable is DAMT which is the donation amount and is non-negative. DONR and DAMT are missing 2007 records from the test set as these will be scored and submitted. While the naming convention of the variables is relatively effective, the data dictionary is a helpful resource. The data set itself has a wide variety of statistics that appear analytically interesting. Table 2 below provides an alphabetic list of the variables, descriptions and data types.

Table 2: Data Dictionary and Data Types

Variable	Description	Type
REG1	Region 1	Binary
REG2	Region 2	Binary
REG3	Region 3	Binary
REG4	Region 4	Binary
HOME	(1 = homeowner, 0 = not a homeowner)	Binary
CHLD	Number of children	Count
HINC	Household income (7 categories)	Count
GENF	Gender (0 = Male, 1 = Female)	Binary
WRAT	Wealth Rating	Count
AVHV	Average Home Value in potential donor's neighborhood in $ thousands	Continuous
INCM	Median Family Income in potential donor's neighborhood in $ thousands	Continuous
INCA	Average Family Income in potential donor's neighborhood in $ thousands	Continuous
PLOW	Percent categorized as “low income” in potential donor's neighborhood	Continuous
NPRO	Lifetime number of promotions received to date	Continuous
TGIF	Dollar amount of lifetime gifts to date	Continuous
LGIF	Dollar amount of largest gift to date	Continuous
RGIF	Dollar amount of most recent gift	Continuous
TDON	Number of months since last donation	Continuous
TLAG	Number of months between first and second gift	Continuous
AGIF	Average dollar amount of gifts to date	Continuous
DONR	Classification Response Variable (1 = Donor, 0 = Non-donor)	Continuous
DAMT	Prediction Response Variable (Donation Amount in $)	Continuous

A listing of the possible predictor variables and theoretical effects is also provided for reference (Table 3). The theoretical effects are particularly important as this logic will be referenced when possible during the examination of candidate variables in the model building process.

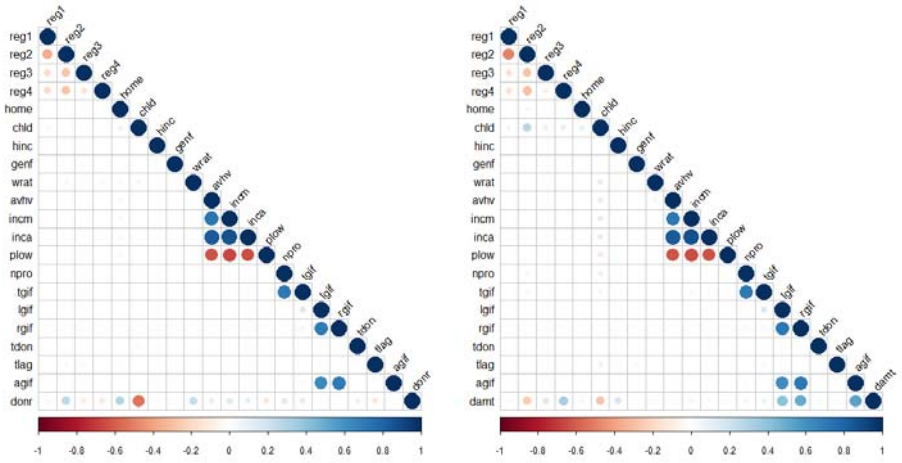
Table 3: Theoretical Effects of Predictor Variables

Variable	Theoretical Effects	Direction
REG1	Unknown effect on probability of donation or dollar amount.	Unknown
REG2	Unknown effect on probability of donation or dollar amount.	Unknown
REG3	Unknown effect on probability of donation or dollar amount.	Unknown
REG4	Unknown effect on probability of donation or dollar amount.	Unknown
HOME	Home owners more likely to donate as they may have the financial means to do so.	Positive
CHLD	Family with more children may not donate or ay donate lower amounts.	Negative
HINC	People with higher incomes might donate and at higher amounts.	Positive
GENF	Unknown effect on probability of donation or dollar amount.	Unknown
WRAT	People at higher wealth levels might donate and at higher amounts.	Positive
AVHV	Higher home values equate to wealth and a propensity to donate more.	Positive
INCM	Wealthier communities likely donate and at higher amounts.	Positive
INCA	Higher family incomes might donate and at higher levels.	Positive
PLOW	Areas with higher % of low income might not donate and donate less.	Negative
NPRO	More promotions lead to more donations but unknown how much.	Positive
TGIF	Likely to donate again but unsure at lower or higher levels than past amounts.	Positive
LGIF	Likely to donate again but unsure at lower or higher levels than past amounts.	Positive
RGIF	Likely to donate again but unsure at lower or higher levels than past amounts.	Positive
TDON	Longer periods since past donations increases propensity to donate again	Positive
TLAG	Shorter time between gifts might lead to more donations but amount unsure	Positive
AGIF	Higher average gifts could mean donating often and at higher amounts.	Positive

Correlations with DONR. The correlations of all the possible independent variables with the responses are examined using the entire data set. These results are provided in Exhibit 1 below. The color intensity and size of the circles are proportional to the correlation coefficients. HOME, REG2 and WRAT are mildly positively correlated with DONR ; 2 of these are in the predicted direction. Intuitively, the relationships with HOME and WRAT makes sense as families who are homeowners and have higher wealth ratings likely have the financial flexibility to donate to charities. REG2's relationship with DONR is difficult to interpret as not much information was given about the variable. Conversely, CHLD is strongly negatively correlated with DONR. This is also logical and in the predicted direction.

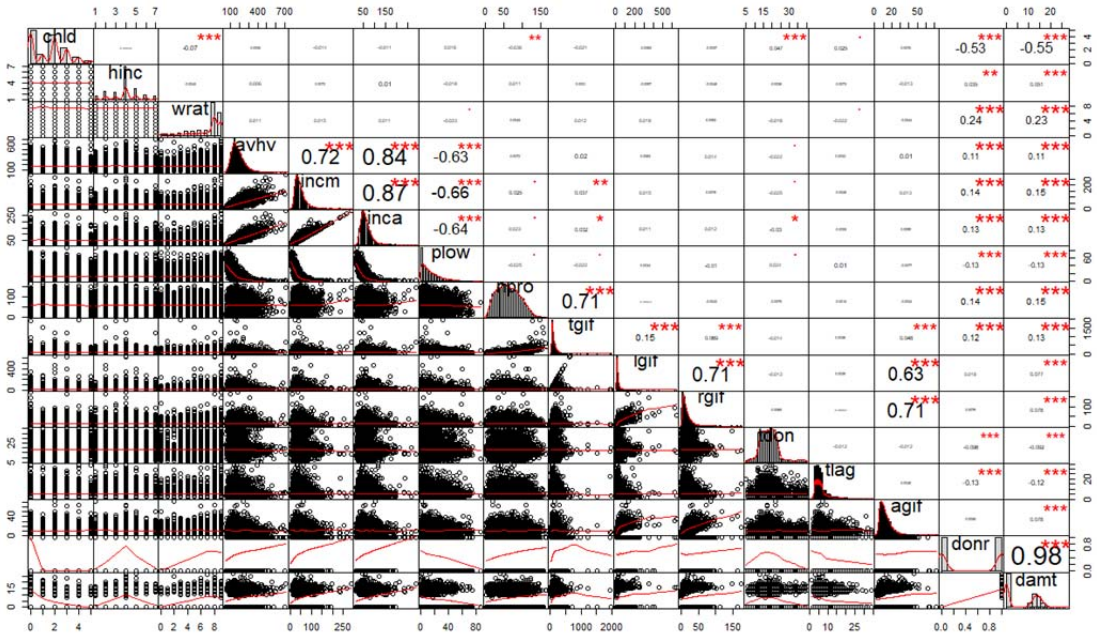
The magnitude of these finds is somewhat disappointing as we hope for a correlation of +0.5 (-0.5) or more (less) with a response variable. There also appears to be negative correlations between the region variables. PLOW has strong negative relationships with AVHV, INCM and INCA (which happen to be strongly positively correlated with each other). This suggests that neighborhoods with higher average home values and income metrics are less likely to be classified as low income. There are several other less pronounced relationships that are not mentioned but will be monitored as the analysis evolves.

Exhibit 1: Correlations with DONR (Left) and DAMT (Right)



Correlations with DAMT. Exhibit 1 also provides the correlations of all possible predictors with DAMT, or the predicted donation amount. Note that this visual represents the full data set only when a donation has been made (DONR = 1). Most variables were in the predicted direction. Metrics pertaining to the average dollar amount (AGIF), amount of most recent gift (RGIF) and amount of largest gift to date (LGIF) all have positive correlations with DAMT greater than 0.4. Number of children (CHLD) and Region 2 (REG2) have the strongest negative correlations (<-0.2). Again, it is ideal to have more polarized correlations between the predictors and response (i.e. closer to -1 or 1). At this juncture the number of children looks to be indicative of a person's propensity to donate, and if they do donate, it is a smaller gift amount if they have more children. There also appears to be similar relationships between the independent variables, which is discussed in the following section.

Exhibit 2: Intercorrelations and Additional Diagnostics of Variables

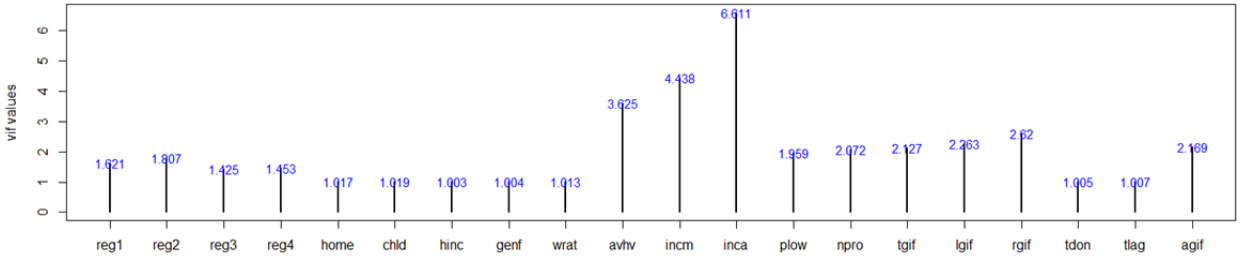


Variable Interactions and Multicollinearity. Exhibit 2 above, produced by the *PerformanceAnalytics* package on the full data set, displays the distribution of each variable on the diagonal, bivariate scatter plots with a fitted line on the bottom diagonal and the value of the correlation well as the significance levels as stars (for example "***" represents a p-value of 0.001). The binary variables did have statistically significant levels of correlation among each other, albeit small effects, but little among the other variable types and thus were removed from the exhibit. The aforementioned relationships between PLOW, AVHV, INCM and INCA is corroborated, and AGIF is also meaningfully and statistically correlated with NPRO, RGIF and LGIF. The first group of variables seem to be associated with a donor's wealth status while the second encapsulates how a donor has historically given to the charity. Using the *usdm* package, the data is checked for multicollinearity via variance inflation factors (VIFs). For reference, a VIF of 10 and above would elicit concern. These values are presented in Exhibit 3 and suggest that multicollinearity is not an issue.

Understanding an interaction effect in a regression model is usually fairly difficult. Using the *jtools* package, interactions are quickly fitted and the coefficients and significance observed with the raw training data. When combined into logistic and multiple regression models, none of the variables identified above showed significant effect of the responses. However, following additional iterations with other potential predictors, there was a significant effect on DONR and DAMT in the presence of significant interaction between CHLD and HOME.

Separately, the results from Exhibit 2 are encouraging as there appears to be recognizable linear relationships between several of the independent variables and DAMT. These variables include CHLD, WRAT, INCM, INCA, PLOW, NPRO, LGIF, RGIF and AGIF. These fits could possibly be improved through removal of the "0" variables as they appear zero inflated in their current form and possible transformation.

Exhibit 3: Variance Inflation Factors



Outliers and Variable Transformations. Table 1 and Exhibit 2 have already suggested the presence of outliers in the potential predictor variables as well as several candidates for transformation. Exhibits 4 and 5 display box and Q-Q plots of the independent variables in the training data set. Please note discussion of how the outliers and transformations are handled is reserved for the subsequent section on data preparation.

Outliers. The majority of variable distributions are right (positively) skewed except for WRAT. Variables that appear to have possible outliers are AVHV, INCM, INCA, PLOW, NPRO, TGIF, LGIFTDON and TLAG. There could be a number reasons for the presence outliers. First, there are data integrity issues when dealing with publicly available data sets. We also know the original proprietors of the data altered and scaled the data differently than what was initially measured. Additionally, having not collected these data ourselves we are forced to rely on the original practitioners in terms of collection techniques and accuracy.

Variable Transformations. In terms variable transformations, several of the variables are candidates for transformation given the insight Table 1 and Exhibit 2 provides into their skewness and kurtosis. Many of the Q-Q plots in Exhibit 4 and 5 show systematic patterns of progressive departures from normality. Given that the data is comprised of counts and measurements that cannot be negative, re-expression will likely benefit the variables.

Exhibit 4: Numerical Variable Outlier and Normality Analysis

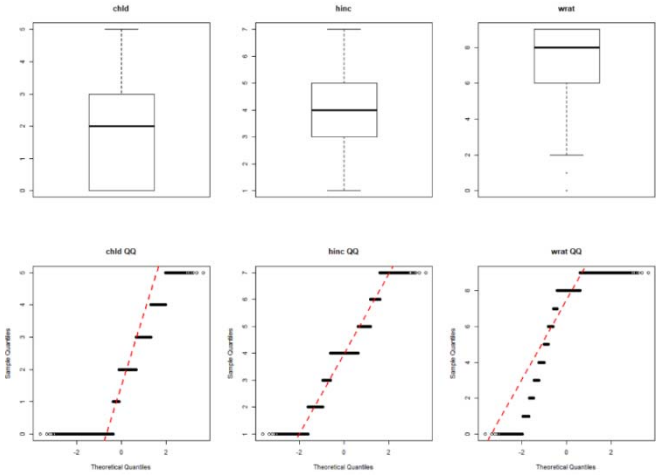
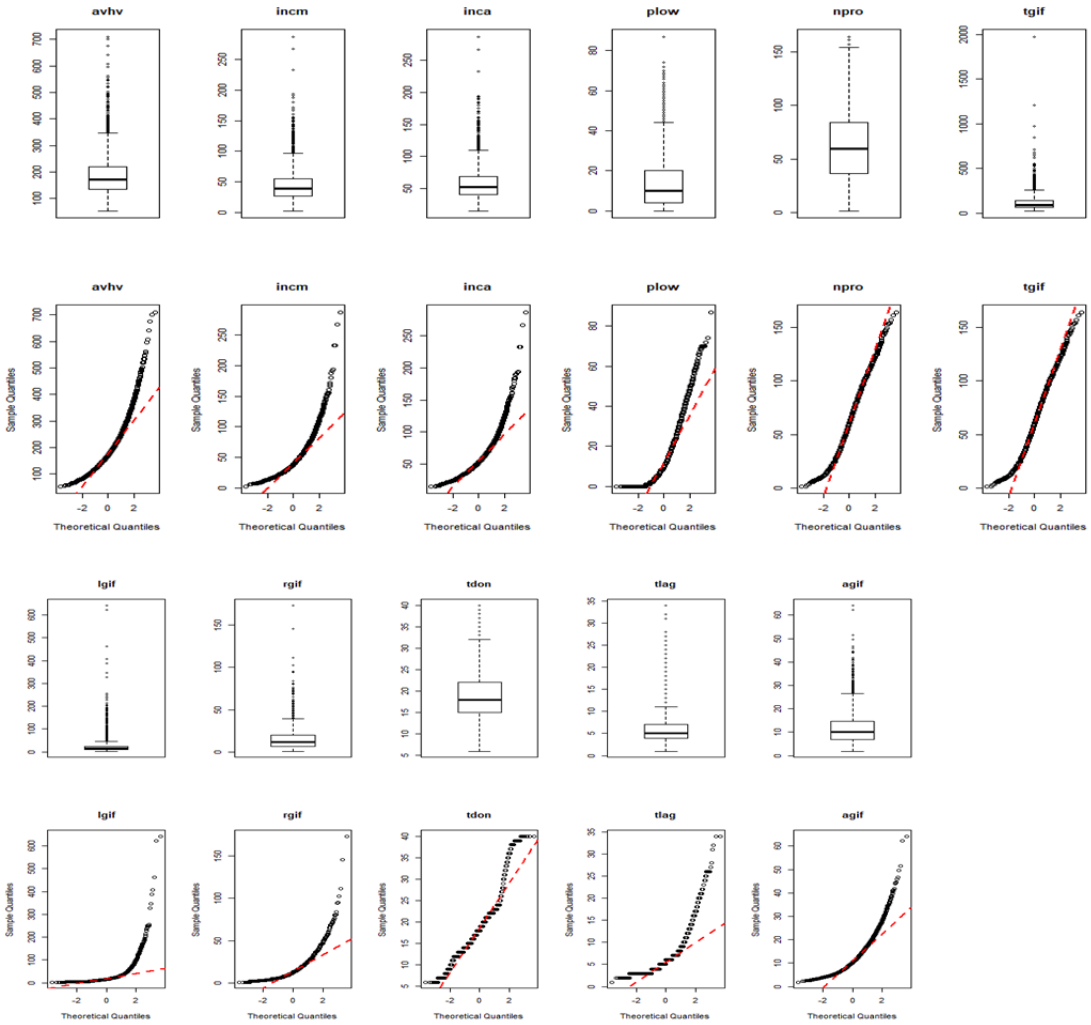


Exhibit 5: Categorical Variable Outlier and Normality Analysis



Classification Model DONR: Identifying Variables of Interest. In the subsequent section, we further identify variables of interest to predict a person's propensity to donate to charity.

Binary Variables. Using the *xtab* function, we view the binary variables by DONR and examine which categories tend to donate more frequently. The objective is to corroborate the findings from the correlation analysis and possibly identify additional predictors where a meaningful difference in values could signal predictive power. Table 4 provides this output. Variables that are bolded appear analytically interesting. To summarize, there is a greater propensity for someone to donate if they live in region 2, are homeowners and/or are male.

Table 4: Binary Variables – Frequency of Donation

Variable		No Donate	Donate	Difference
REG1				
	No	1627	1541	86
	Yes	362	454	-92
REG2				
	No	**1553**	**1092**	**461**
	Yes	**436**	**903**	**-467**
REG3				
	No	1675	1817	-142
	Yes	314	178	136
REG4				
	No	1635	1812	-177
	Yes	354	183	171
HOME				
	No	**417**	**48**	**369**
	Yes	**1572**	**1947**	**-375**
GENDER				
	Female	**769**	**805**	**-36**
	Male	**1220**	**1190**	**30**

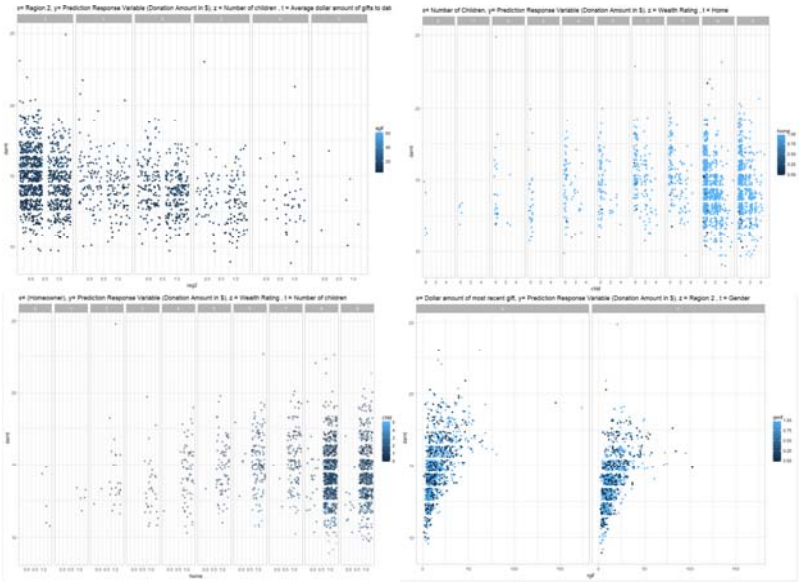
Mean Analysis for Count and Continuous Variables. Using the *tapply* function grouped by DONR, the data set is decomposed further to compare the mean values for all numeric variables by 0 and 1 (Table 5). The objective is similar to the preceding binary variable frequency analysis where differences in average values could suggest predictive power. While the relative change needs to be considered with regard to the respective variables, several insights can be extracted from the table. There appears to noteworthy differences in CHLD, AVHV, NPRO and TGIF. A given individual will donate more when they have fewer children, a higher average home value (a wealth indicator), received more promotions to date and have donated more cumulatively over time. These variables of interest are bolded in the table below.

Table 5: Mean Analysis of Numeric Variables by DONR

	No Donate 0	Donate 1	Difference
Count Variables			
CHLD	**2.33**	**0.83**	**1.50**
HINC	3.91	3.98	-0.08
WRAT	6.47	7.63	-1.16
Continuous Variables			
AVHV	**176.28**	**194.06**	**-17.78**
INCM	40.30	48.26	-7.96
INCA	53.61	60.65	-7.04
PLOW	15.61	11.86	3.75
NPRO	**57.51**	**65.74**	**-8.23**
TGIF	**106.83**	**126.63**	**-19.80**
LGIF	22.39	23.99	-1.61
RGIF	15.37	15.73	-0.36
TDON	19.35	18.28	1.07
TLAG	6.81	5.79	1.02
AGIF	11.60	11.72	-0.12

Prediction Model DAMT: Identifying Variables of Interest. Exhibit 2 earlier illustrated that most of the continuous variables have somewhat of a linear relationship with DAMT. Now we examine categorical and count variables that might be predictive of DAMT. Beginning with the variables that are highly correlated with DAMT (CHLD, REG2), several different visuals are presented below in Exhibit 5. Note that the training data set has been partitioned to only include records where DONR is equal to 1. Of the highly correlated variables with DAMT, the number of children was consistently indicative of donation amount and, specifically, the more children someone has the lowers the donation amount is. Other findings include that males give outsized (i.e. extreme) donations more than females, homeowners donate more frequently at higher wealth levels (reduced by number of children however) and most people in this sample are homeowners.

Exhibit 5: Scatterplots for Variables with Highest Correlation to DAMT



Variable Selection. Numerous techniques for variable selection such as k-folds cross validation, best subset, to name a few, are used in the proceeding sections for model construction. The *randomForest* package is used here observe which training variables are considered the most important by the Mean Decrease Gini metric. The higher this value, the more important the variable is in predicting the responses. Exhibits 6 and 7 confirm that CHLD could be an important predictor for both models. Surprisingly, home ownership and the regional categories ranked poorly in each evaluation. From this analysis and the initial EDA, I posit that the number of children and various wealth indicators will carry predictive power in the classification models while the prediction model will be driven more by numerical variables associated with gifting amounts over time.

Exhibit 6: Classification Model Variable Importance (DONR)

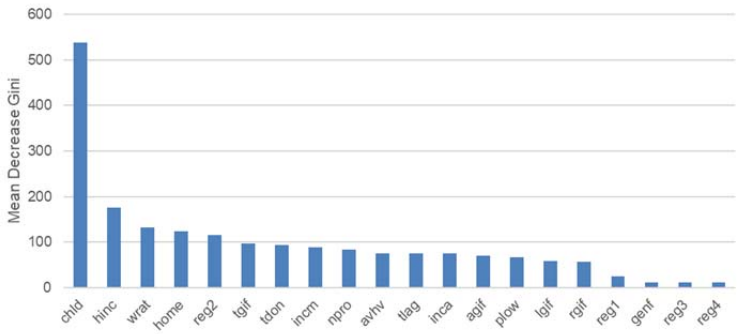
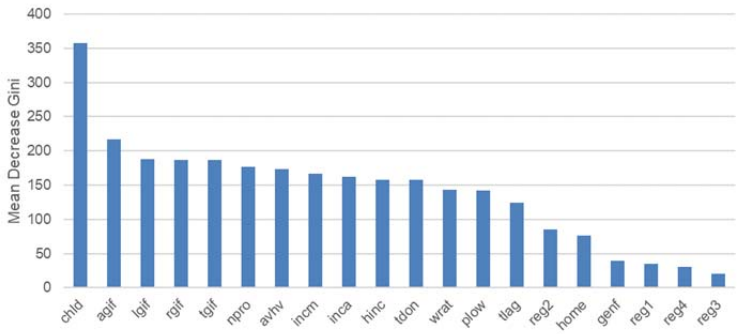


Exhibit 7: Predictive Model Variable Importance (DAMT)



The EDA uncovered several interesting elements in the data set. Overall, we have a better understanding of the dynamics underpinning the data set as well as identified several variables for inclusion in the classification and prediction models when attempting to improve performance via manual variable reduction. Below is a summary of the variables that will be used in the reduced models.

- ***Classification Model:***
  - CHLD_IMP*HOME
  - HINC
  - TGIF_IMP
  - TDON_IMP
  - NPRO_IMP

- ***Predictive Model:***
  - CHLD_IMP*HOME
  - AGIF_IMP
  - LGIF_IMP
  - RGIF_IMP
  - NPRO_IMP
  - PLOW_IMP

DATA PREPARATION. In this section, the data is prepared by changing the original values and replacing them with new modified variables. Per instructions from the course instructor, the 2 response variables are left unchanged.

Outlier Resolution. In attempt to create a better fit for the models, we first modify the data set to eliminate possible outliers as they can significantly influence regression results. To accomplish this, we classify outlier values that are greater than the third quartile or less than the first quartile and replace them with the interquartile range (IQR) multiplied by 1.5. Most of the variables had high value outliers which were replaced. The new variables are labeled "_IMP" meaning it has been imputed and replaced the original value.

Variable Transformation Resolution. Each count and continuous variable was analyzed for departures from normality and then evaluated using a series of different transformations to find the normality improvement. Table 6 provides a summary of these re-expressions and their impact on the independent variables. Overall, the distributions appear much more normally distributed.

Table 6: Outlier and Variable Transformation Resolutions

Variable	Re-Expression	Original Skewness	New Skewness	Skewness Change	Original Kurt.	New Kurt.	Kurt. Change
CHLD	Square Root	0.27	-0.47	*0.74*	-0.80	-1.20	*0.40*
WRAT	Squared	-1.35	-0.78	*0.57*	0.79	-0.72	*1.51*
AVHV	Log	1.54	-0.09	*-1.63*	4.49	-0.30	*4.79*
INCM	Log	2.05	1.14	*-0.91*	8.31	3.09	*5.22*
INCA	Log	1.94	-0.09	*-2.03*	7.87	-0.30	*8.17*
PLOW	Sin	1.36	-0.18	*-1.54*	1.89	-1.38	*3.27*
NPRO	Sin	0.31	0	*-0.31*	-0.62	-1.50	*0.88*
TGIF	Square Root	6.55	0.62	*-5.93*	107.52	-0.51	*108.03*
LGIF	Log	7.81	-0.09	*-7.9*	110.38	-0.69	*111.07*
RGIF	Log	2.63	0	*-2.63*	13.92	-0.19	*14.11*
TDON	Log	1.1	-0.4	*-1.5*	2.12	0.90	*1.22*
TLAG	Log	2.42	0.09	*-2.33*	8.41	-0.71	*9.12*
AGIF	Log	1.78	-0.1	*-1.88*	6.02	-0.60	*6.62*

Missing Data Resolution. There were no missing data points in the train and validation data sets.

The resultant structure of the cleaned data set (not shown) has a total of 24 variables overall with no missing values, 2 imputed count variables and 11 imputed continuous variables to remedy outliers and improve fit with the response. Outlier deletion and imputation are intended to improve the robustness of models and can have powerful effects on fit. While the effects are sometimes not in the desired direction, these analytical activities can be beneficial if they improve the relationship between two or more variables. Finally, the cleaned variables are standardized to avoid possible issues that could arise from advanced modeling techniques and then partitioned into training (3984), validation (2018) and test (2007) sets using the PART flag.

MODEL PRODUCTION AND SELECTION. According to the data provider's recent mailing records, the typical overall response rate to a direct marketing campaign is 10%. Out of those who respond (donate) to the mailing, the average donation is $14.50. Each mailing costs $2.00 to produce and send; the mailing includes a gift of personalized address labels and assortment of cards and envelopes. It is not cost-effective to mail everyone because the expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55. The first objective is to 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.

Classification Model. We begin by fitting series classification models to predict the chance a potential donor will respond to a marketing campaign. The modeling techniques include logistic regression, logistic regression general additive model (GAM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN), decision trees, bagging, boosting and support vector classification. All candidate models are fit on the training data and then evaluated using the validation data. Each candidate model is fit with all modified variables as a "base case". Some then used the reduced variable sets identified during the EDA in hope of improving performance. Maximum profit is the ultimate evaluation criteria and the final selected classification model will be used to classify DONR responses in the test data set.

A total of 19 different models were generated, however, for the sake of brevity only the top performing model per technique is presented in the statistical summary in Table 7 below. The models that included all the modified variables or used some form of automated selection technique performed better than the manually reduced variable set. The logistic and logistic GAM models initially performed very similarly so a natural spline was added to CHLD in the logistic GAM model for experimentation as it is a potentially important variable. The tree-based models appeared to perform worse when using cross-validation for pruning while the KNN improved meaningfully between K=1 and K=10. From a comparative statistic standpoint, the tree-based models produced similar true positive and true negative rates over the validation set and outperformed in terms of accuracy. KNN performed better than the logistic, LDA and QDA models, indicating that the decision boundary might be non-linear. The logistic and logist GAM models' strengths seem to be their true negative values and relative accuracy. Examining the area under the curve (AUC) we see that each model is performing between 83% to 89% coverage. At this juncture, any additional iterations will likely only generate incremental improvement.

Table 7: Classification Model Comparative Statistics

	Log.	Log. GAM	LDA	QDA	KNN	Tree	Bag	Boost	SVC
Accuracy	0.79	0.80	0.75	0.77	0.82	0.85	0.89	0.83	0.84
Kappa	0.58	0.61	0.50	0.55	0.63	0.70	0.77	0.67	0.69
Sensitivity	0.97	0.97	0.99	0.98	0.93	0.93	0.90	0.85	0.86
Specificity	0.61	0.64	0.51	0.58	0.71	0.77	0.87	0.82	0.83
Pos Pred Value	0.71	0.72	0.67	0.69	0.76	0.80	0.87	0.82	0.83
Neg Pred Value	0.96	0.96	0.98	0.96	0.91	0.92	0.90	0.85	0.86
Prevalence	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Detection Rate	0.48	0.48	0.49	0.48	0.46	0.46	0.45	0.42	0.43
Detection Prev.	0.68	0.67	0.74	0.70	0.61	0.58	0.51	0.51	0.51
Balanced Acc.	0.79	0.80	0.75	0.78	0.82	0.85	0.89	0.83	0.84
AUC	83.53%	84.19%	82.13%	82.84%	83.22%	85.77%	88.84%	83.33%	84.43%

Table 8 provides the results of the classification models in terms of number of mailings that maximize profit in the validation sample. The logistic GAM was the best performing model while KNN, support vector classifiers, boosting and bagging were the worst. Overall, the recommendation is to select the base logistic GAM model with the natural spline as it earned the largest predicted profit in the validation sample. The natural spline is a flexible way of fitting non-linear models and learning the non-linear interactions from the data. It should be noted that none of the models outperformed a raw, untransformed logistic model (not shown) but came very close. Due to oversampling in the training and validation sets, the test data must be adjusted. The response rate in the test sample has the more typical 10% response rate and is modified to reflect this assumption. Based on the adjusted model, the organization should mail to the 363 highest posterior probabilities.

Table 8: Classification Model Results

Model Technique	Mailings	Maximum
Logistic	1,369	$11,370.50
Logistic GAM	***1,434***	***$11,408.00***
LDA	1,484	$11,343.50
QDA	1,410	$11,346.50
KNN	1,227	$10,987.50
Decision Tree	1,165	$11,140.50
Bagging	1,035	$11,038.00
Boosting	1,030	$10,207.00
Support Vector Classifier	1,038	$10,408.50

Prediction Model. As the second part of the exercise, a prediction model is built to predict the expected gift amounts from donors (DAMT). The data for this will consist of the records for donors only. Similar to the prior section on classification modeling, a series of candidate models are fit using the training data and then evaluated using the validation data. The techniques used to predict the gift amount include ordinary least squares (OLS), best subset selection, principal component regression (PCR), partial least squares (PLS), ridge regression, lasso, decision tree, bagging, and boosting. The reduced variable sets identified during the EDA are only included in 3 of the models while cross-validation is used for the other models involving variable or component selection. The minimum mean prediction error is the evaluation criteria for the best prediction model.

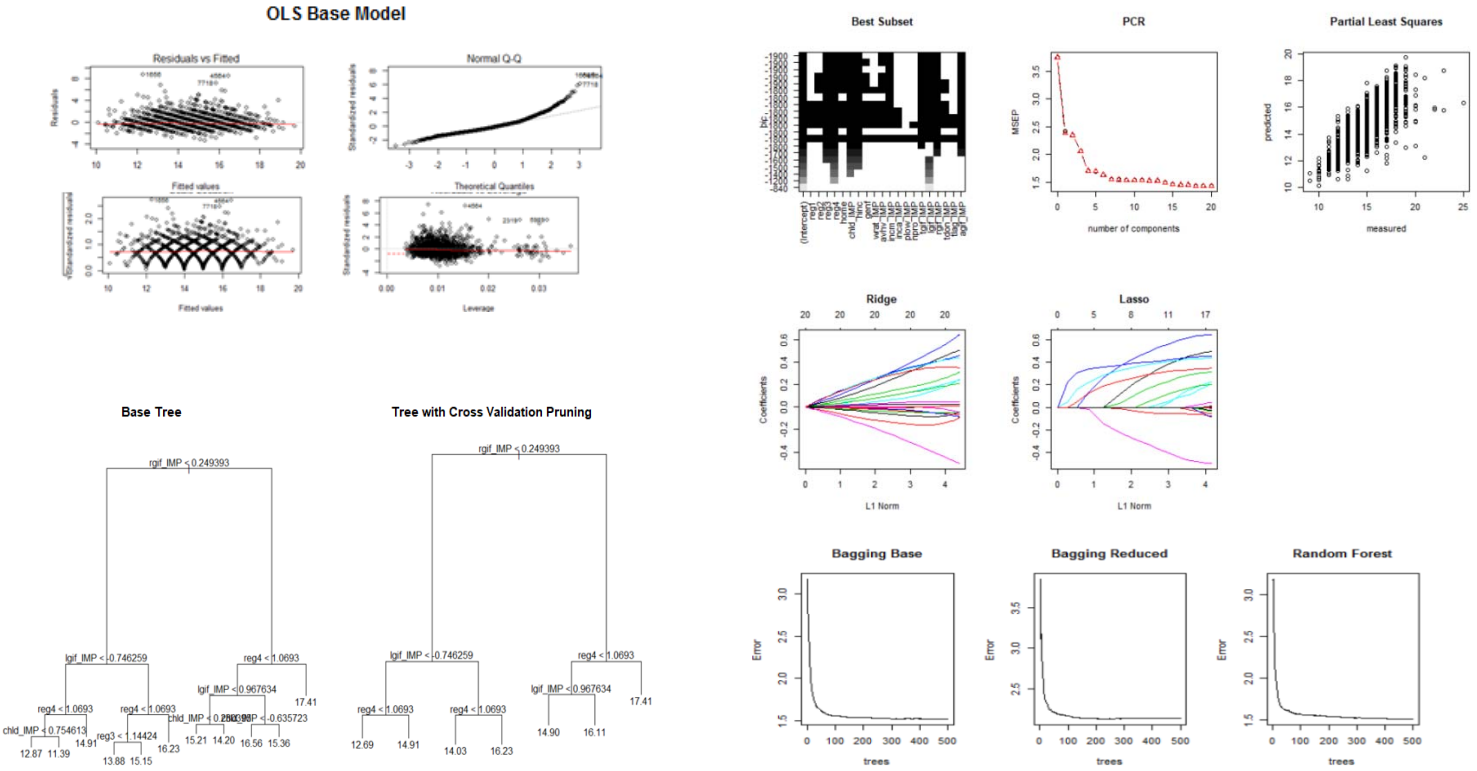
A total of 16 models were generated and their results are presented in Table 9 below. The models that did not use the reduced variable set from the EDA performed better once again. Cross-validation was used in several models and appears effective in improving performance. The tree-based models generally performed worse with pruning via cross-validation not adding much value. Exhibit 8 provides a series of diagnostic plots for the respective models. With the OLS model the scatterplot for the residual values appears symmetrically clustered towards the middle of the plot, signaling possible issues with the model, however, the Q-Q plot suggests a relatively normal distribution. The best subset model included 14 predictors, including several continuous and categorical as well as regional variables, which is surprising as the regional variables were not identified earlier as useful. The PCR plot suggests there is an elbow at 5 components and not much benefit beyond, however, it was found with the validation set that when including 20 components there was a meaningful improvement. The opposite was found with PLS regression, where including more components hurt. The ridge and lasso models performed quite similarly, evident in their respective coefficient plots. The decision tree was not enhanced through pruning of the terminal nodes (best at 5), though interestingly the splits mainly consist of variables associated with gifting (dollar amount gifted, etc.) and regional categories originally thought to be insignificant. Lastly, the bagging, boosting and random forest models show there was not much benefit to error reduction beyond about 100 trees. Interestingly, REG4, CHLD_IMP, RGIF_IMP appear to be the more important predictors (not shown) in the tree based models as well as lasso as they enter the model earlier.

Table 9: Prediction Model Results

Prediction Model	Mean Prediction Error	Standard Error
OLS Base	1.638	0.163
OLS Reduced	2.346	0.197
Best Subset w/ CV	1.643	0.163
PCR 5 Comp.	1.859	0.169
PCR 20 Comp.	1.638	0.163
PLS 2 Comp.	***1.636***	***0.162***
PLS 15 Comp.	1.638	0.163
Ridge w/ CV	1.642	0.164
Lasso w/ CV	1.639	0.163
Decision Tree	2.241	0.192
Decision Tree Pruned	2.379	0.187
Bagging Base	1.755	0.178
Bagging Reduced w/Tree	2.430	0.202
Boosting Base	2.631	0.217
Boosting Reduced w/Tree	2.338	0.197
Random Forest	1.706	0.175

The recommendation for the prediction model is to select the PLS model with 2 components as it has the lowest mean prediction error rate. The logistical GAM classification and PLS with 2 components forecast a profit of $5,229.75 from mailing to 304 potential donors.

Exhibit 8: Prediction Model Diagnostic Plots



CONCLUSION

Building sound machine learning models is a cornerstone of predictive analytics and to creating a culture of proactive data-driven decision making. For this exercise, a historical data set was used to construct classification models to predict the propensity of individuals to donate and, in turn, how much is donated. The potential value of enhanced predictive forecasting for charitable organizations is evident. If a charity can predict what characteristics of past donors increase the likelihood and magnitude of future gifts, then the organization will be able to adjust their marketing efforts to maximize profits by decreasing costs through focused targeting.

The data set for this analysis was diverse and the selected models are not flawless but do provide the best results within the constructs of the exercise and modifications to the original data set. The logistic GAM classification and partial least squares with 2 components showed the best performance and given that none of the manually selected reduced models were chosen it appears that the value brought to this exercise by the analyst was the cleaning and transformation of the variables to improve fit. The inclusion of the interaction term in the manually reduced sets could have also been problematic as these variables introduce new levels of complexity to an analysis. A thorough vetting of these variables, along with further testing of transformations and non-linear tuning could have improved the performance of the models.