Optimizing Mail Campaigns:

An Analysis of Costs and Earnings for Nonprofit Organization ABC

Jose Faz, Morgan Hellwig, Trisa Nguyen, and Ashley Pita

University of North Texas

ADTA 5230: Advanced Data Analytics II

Dr. Anna George

December 9, 2020

Executive Summary

To improve the cost effectiveness of a nonprofits' mail campaigns, our team analyzed data from their most recent campaign. Using four classification models, we predicted the most likely classification of donor or non-donor in future campaigns. Of the models applied, the Nominal Logistic Regression model yielded the highest accuracy and return on investment. We also applied fourlikely donation amount of those classified as donors. Of the models applied, the least root mean squared error was identified in the Stepwise Linear Regression model. Our recommendation is to apply these models for future campaigns to minimize waste in mailing unlikely donors and maximize their return on investment.

Optimizing Mail Campaigns:

An Analysis of Costs and Earnings for Nonprofit Organization ABC

Introduction

Every day, millions of lives benefit from the work of nonprofit organizations. In the United States alone, there are 1.3 million charitable nonprofits that strive to engage the public in politics, strengthen the economy, and serve individuals in local communities or nationwide (Council of Nonprofits, 2020). Many Americans are not even aware of the numerous benefits nonprofits have had for them. While there may be varying interpretations on what a “nonprofit” organization is, all definitions have one requirement in common: no individual may benefit from the organization’s net earnings. This means that growth of funds will only further the organization’s mission rather than provide financial incentive to shareholders. One way that nonprofits bring in revenue is through donations.

Nonprofits may utilize many approaches to generating donations, from grand social events to a simple telephone campaign. Most methods involve investing a portion of the organization’s funds in anticipation that the donations generated will outweigh them, resulting in a net gain. Unfortunately, not all fundraisers are a success. However, as data analytics have increasingly been utilized in for-profit businesses, nonprofit organizations have also learned that they can leverage data to supplement their decision-making, and ultimately maximize funds for their cause.

Business Understanding

This report details the study performed to improve the cost effectiveness of mail campaigning for Nonprofit Organization ABC, where cost effectiveness is defined as return on investment (ROI = Net Funds Gained / Capital Investment × 100). The mail campaign encourages individuals to donate by sending marketing letters and personalized address labels to prior donors. The cost to produce and send each mailing is $2.00 and the average donation generated is $14.50. Unfortunately, not every household that receives a campaign mailing will respond with a donation. Based on historical data, only an approximate 10% of recipients donate.

The intent of this study is to best answer three questions: 1) which donors are most likely to respond to the mailing campaign with a new donation, 2) what are the predicted donation amounts of said donors, and 3) how can this information be used to maximize return on investment? The resulting predictions can then be leveraged to optimize the net proceeds for the organization's mission (Donations Generated – Capital Investment = Net Funds Gained), and ultimately maximize the ROI for the campaign.

Data Understanding

The data that has been made available to our team consists of 8,009 observations from previous donors in the most recent campaign, each consisting of 22 variables. The target variable for classification is DONR (whether a recipient donated) and for prediction is DAMT (the donation amount). The data has been partitioned into three groups: training (3,984 observations, 50% of data set), validation (2,008 observations, 25% of data set), and testing (2,007 observations, 25% of data set). To ensure each group is well represented for the analysis, the training and validation data set has been weighted to be balanced among donors and non-donors. Hence, while donors typically only make up 10% of the population, they are approximately 50% of our training and validation data set. The test data has not been weighted. The variables in the available data are summarized below in Table 1.

Table 1

Data Dictionary of Original Variables Provided

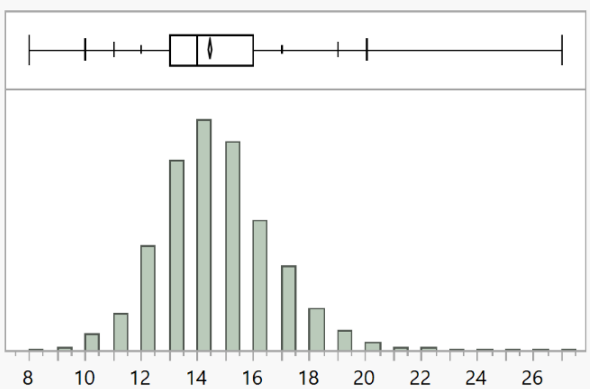
Name	Description	Type
OWN	Donor is a homeowner (1) or not a homeowner (0)	Nominal
KIDS	Number of children the donor has	Continuous
INC	Donor’s household income level (1-7)	Ordinal
SEX	Donor’s is male (0) or female (1)	Nominal
WLTH	Donor’s wealth rating (0-9)	Ordinal
TER1	Donor resides (1) or does not reside (0) in region 1 (of regions 1-5)	Nominal
TER2	Donor resides (1) or does not reside (0) in region 2 (of regions 1-5)	Nominal
TER3	Donor resides (1) or does not reside (0) in region 3 (of regions 1-5)	Nominal
TER4	Donor resides (1) or does not reside (0) in region 4 (of regions 1-5)	Nominal
HV	Average Home Value in potential donor's neighborhood in $ thousands	Continuous
INCMED	Median Family Income in potential donor's neighborhood in $ thousands	Continuous
INCAVG	Average Family Income in potential donor's neighborhood in $ thousands	Continuous
LOW	Percent categorized as “low income” in potential donor's neighborhood	Continuous
NPRO	Lifetime number of promotions received to date	Continuous
GIFDOL	Dollar amount of lifetime gifts to date	Continuous
GIFL	Dollar amount of largest gift to date	Continuous
GIFR	Dollar amount of most recent gift	Continuous
MDON	Number of months since last donation	Continuous
LAG	Number of months between first and second gift	Continuous
GIFA	Average dollar amount of gifts to date	Continuous
DONR	Latest campaign resulting in donation (1) or no donation (0)	Nominal
DAMT	Donation Amount in $	Continuous
PART	Partition of data (training, validation, test)	Nominal

An initial review of the variables revealed that the average donation amount for the data set was $14.45 with a standard deviation of $1.98 (Figure 1). While the overall range of these donations was $19.00 (minimum $8.00, maximum $27.00), 80% were between $12.00 and $17.00. A slight positive skew is made evident by the histogram, indicating the possibility of high donation outliers.

Figure 1

JMP Output - Distribution Histogram, Box Plot, and Summary Statistics for DAMT

damt



Quantiles

100.0%	maximum	27
99.5%		20.025
97.5%		19
90.0%		17
75.0%	quartile	16
50.0%	median	14
25.0%	quartile	13
10.0%		12
2.5%		11
0.5%		10
0.0%	minimum	8

Summary Statistics

Mean	14.451904
Std Dev	1.9848721
Std Err Mean	0.0362749
Upper 95% Mean	14.52303
Lower 95% Mean	14.380777
N	2994

Variable OWN showed a noteworthy difference in response rates between homeowners and non-homeowners (homeowners responding nearly six times as often), as did INCMED (response rates significantly decreasing towards highest and lowest income areas). The variable KIDS also indicated that likelihood of response tended to decrease as the number of donor's children increased. Gender appeared to have no noteworthy relationship with either response rate or donation amount, however, the dollar amount of gifts to date (GIFA) seems to have a direct correlation with donation amount. These observations, as well as others, were frequently revisited when evaluating both the classification and predictions models. A complete overview of the descriptive statistics by variable is available in Appendix A of this report.

Methods and Analysis

Data Preparation

In preparation for applying classification and prediction models to the data, we updated the data and model types of each variable in the raw dataset to ensure consistency with the data, as shown in Table 1. Next, we identified seven variables with monetary values: HV, INCMED, INCAVG, GIFDOL, GIFL, GIFR, and GIFA. Of these variables, HV, INCMED, and INCAVG were expressed in thousands and the remaining variables were expressed in dollars. To normalize the monetary units used, we transformed the units of the variables expressed in thousands to dollars. These new variables were coded as HV_DOLLARS, INCMED_DOLLARS, and INCAVG_DOLLARS. We then added a binary nominal variable for whether a potential donor had kids or not. While the KIDS variable outlines the number of children a potential donor has, we believed it to be potentially valuable to classify individuals as parents or not parents. This new variable was coded as KIDSB.

Prior to running descriptive statistics on the remaining data, we sorted and reviewed the monetary variables relating to recent gifts: GIFDOL, GIFL, GIFR and GIFA. In reviewing this data, we discovered various discrepancies in these variables. In various situations where GIFL and GIFR were listed as different values, and thus suggesting they were separate gifts, the value of GIFDOL was lower than GIFL and GIFR combined. An example of this situation is ID 4488. We also discovered potential donors with different GIFL and GIFR values, again suggesting separate gifts, with a GIFDOL value the same as GIFR. An example of this situation is ID 364. In addition, we discovered cases were GIFR exceeded GIFL, such as with ID 2858, and cases where GIFDOL was equivalent to GIFR but the LAG and GIFA columns suggested the potential donor has given at least twice before, such as with ID 6877. There was also one case in ID 7759 where GIFA exceeded GIFDOL. Because of these observations and a lack of confidence in the GIFDOL, GIFL and GIFR variables, we decided to exclude these from our analysis and only utilize GIFA instead. This decision was supported by creating a correlation matrix between DAMT, GIFDOL, GIFL, GIFR and GIFA for donors only and discovering GIFA has the strongest relationship with DAMT as suggested by a correlation of 0.5332 as shown in Table 2.

Table 2

Correlation Matrix of DAMT, GIFDOL, GIFL, GIFR, and GIFA for Donors

	gifdol	**gifl**	**gifr**	**gifa**	**damt**
gifdol	1.0000	0.1519	0.0497	0.0205	0.0900
gifl	0.1519	1.0000	0.7126	0.6199	0.4020
gifr	0.0497	0.7126	1.0000	0.7108	0.5096
gifa	0.0205	0.6199	0.7108	1.0000	0.5332
damt	0.0900	0.4020	0.5096	0.5332	1.0000

Note. The correlations are estimated by Row-wise method.

To continue analyzing our data, we ran a correlation matrix for all continuous variables, excluding KIDS, for the entire dataset. In reviewing this data, shown in Table 3, we identified a very strong relationship between INCMED_DOLLARS and INCAVG_DOLLARS. Because the median is less influenced by outliers or skewed data, we chose to utilize INCMED_DOLLARS and exclude INCAVG_DOLLARS in further analysis. Based on the correlation matrix, we also made the decision to exclude HV_DOLLARS and LOW in models that assume no multicollinearity due to the strong correlations of these values with INCMED_DOLLARS. In the case of both HV_DOLLARS and LOW, we believed INCMED_DOLLARS to be a more valuable datapoint for our analysis.

Table 3

Correlation Matrix of Numeric Variables

	hv_dollars	incmed_dollars	incavg_dollars	low	npro	mdon	lag	gifa
hv_dollars	1.0000	0.7262	0.8401	-0.6309	-0.0033	-0.0195	0.0134	0.0072
incmed_dollars	0.7262	1.0000	0.8701	-0.6552	0.0228	-0.0226	-0.0041	0.0134
incavg_dollars	0.8401	0.8701	1.0000	-0.6379	0.0154	-0.0278	0.0033	0.0102
low	-0.6309	-0.6552	-0.6379	1.0000	-0.0206	0.0231	0.0039	-0.0116
npro	-0.0033	0.0228	0.0154	-0.0206	1.0000	-0.0059	0.0088	-0.0056
mdon	-0.0195	-0.0226	-0.0278	0.0231	-0.0059	1.0000	-0.0062	-0.0109
lag	0.0134	-0.0041	0.0033	0.0039	0.0088	-0.0062	1.0000	0.0050
gifa	0.0072	0.0134	0.0102	-0.0116	-0.0056	-0.0109	0.0050	1.0000

Note. The correlations are estimated by Row-wise method.

Prior to running any models, we also reviewed the distributions of our continuous variables, excluding KIDS. In anticipation of running models that assume normally distributed variables, we wanted to determine in any transformations were necessary to meet this assumption. The output in JMP, shown in Figure 2, revealed a positive skew in almost all variables. NPRO was the most normally distributed variable. To account for the positive skews, we created log-transformed variables of each numeric value as shown in Figure 3. The log transformation corrected for the positive skews in the original data, particularly with variables HV_DOLLARS, INCMED_DOLLARS, LOW and GIFA. At this point, we prepared our dataset for the application of classification and prediction models by excluding the test data from analysis. All variables added during data preparation are listed in Table 4.

Figure 2

JMP Output - Histograms and Normal Quartile Plots for Numeric Variables

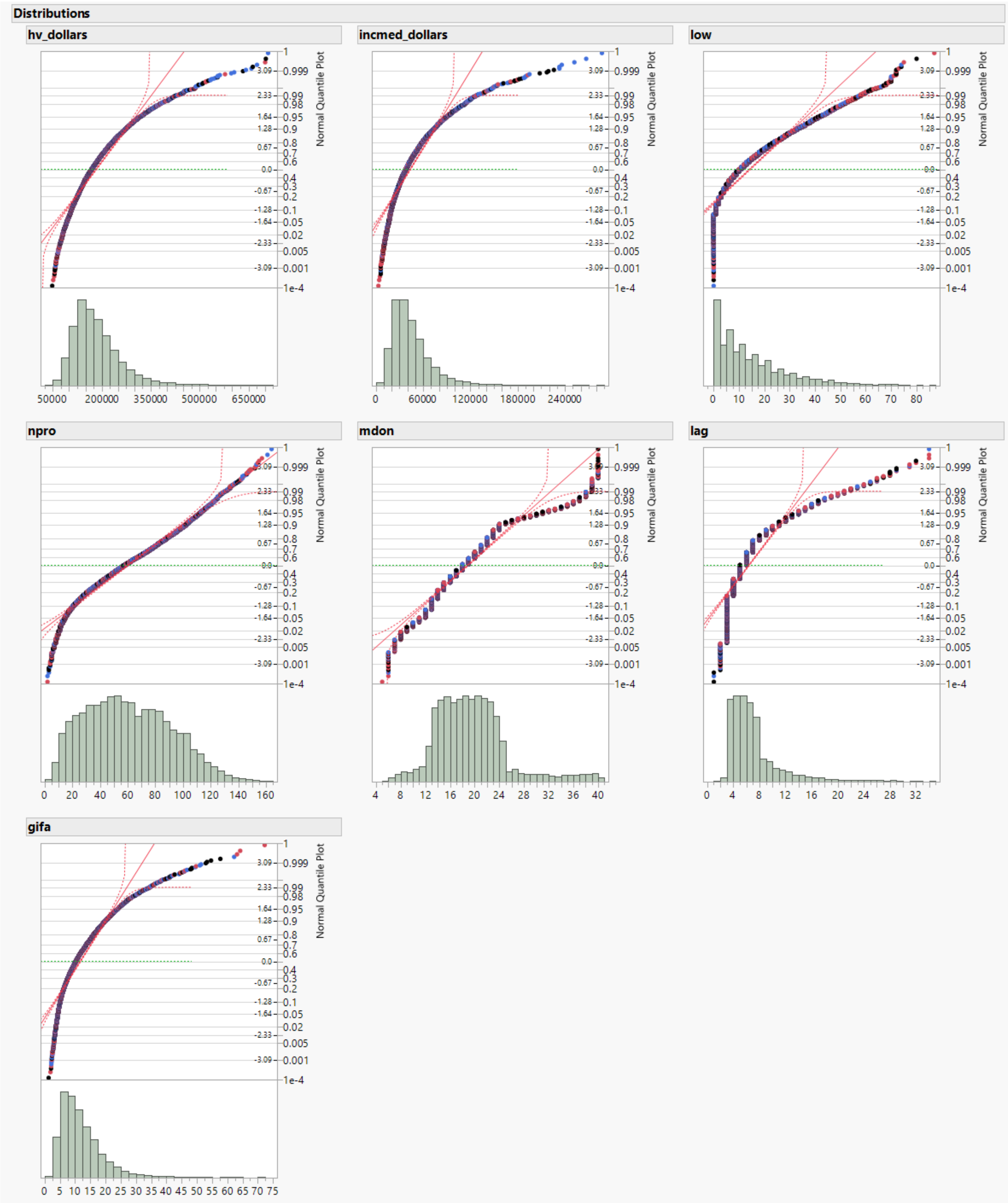


Figure 3

JMP Output - Histograms and Normal Quartile Plots for Log Transformed Numeric Variables

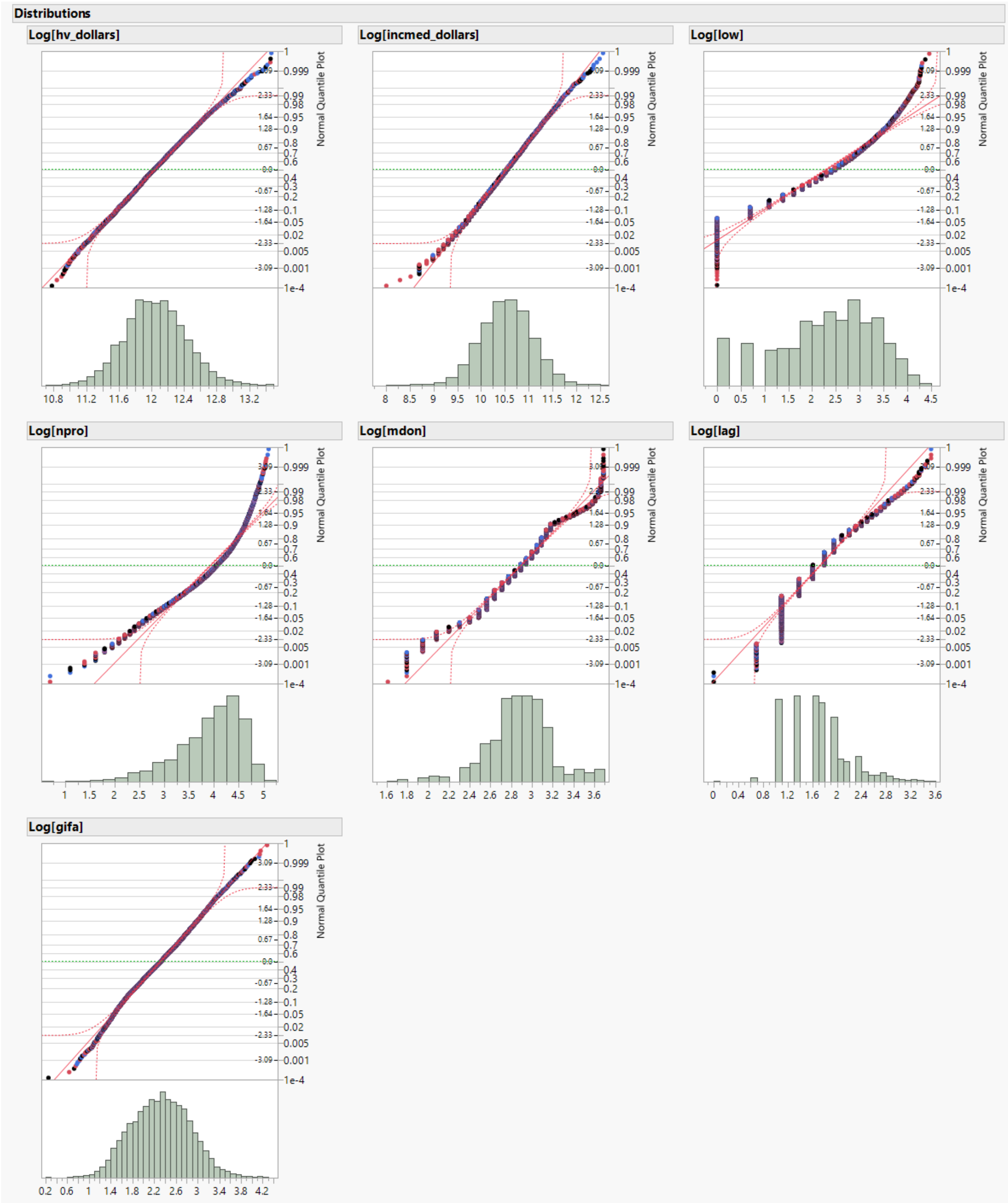


Table 4

Data Dictionary of Variable Added During Date Preparation

Name	**Description**	**Type**
KIDSB	Donor has kids (1) or does not have kids (0)	Nominal
HV_DOLLARS	Average Home Value in potential donor's neighborhood in dollars	Continuous
INCMED_DOLLARS	Median Family Income in potential donor's neighborhood in dollars	Continuous
INCAVG_DOLLARS	Average Family Income in potential donor's neighborhood in dollars	Continuous
LOG[HV_DOLLARS]	Log Transformation of HV_DOLLARS	Continuous
LOG[INCMED_DOLLARS]	Log Transformation of INCMED_DOLLARS	Continuous
LOG[LOW]	Log Transformation of LOW	Continuous
LOG[NPRO]	Log Transformation of NPRO	Continuous
LOG[MDON]	Log Transformation of MDON	Continuous
LOG[LAG]	Log Transformation of LAG	Continuous
LOG[GIFA]	Log Transformation of GIFA	Continuous

Modeling

To predict likely donors in our test dataset, we applied four classification models for the DONR variable to the training and validation data and determined the best fit based on the maximum ROI associated with each model. To predict donation amounts, we applied four prediction models for the DAMT variable to only donors in the training and validation data and determined the best fit model by comparing the root mean squared error of each model. The chosen models and model outputs are detailed in the following sections.

Classification Model 1. To classify donors in our training and validation dataset, we started with a classification tree. Classification trees use splits in predicter variables to separate observations into subgroups (Shmueli et. al, 2017). Strengths of this model are the logic is easily interpreted and there are no assumptions of variables, such as normal distributions or multicollinearity. They are also inexpensive models to deploy. One weakness of trees is relationships between variables may be missed due to the reliance on only one predictor variable at a time. To run this model, we utilized all predictive variables except for KIDSB and the log-transformed numeric variables. Because we had validation data available, we were able to utilize JMPs features to run the model and not need to further prune. The final tree utilized 31 splits. The JMP fit details output is shown in Figure 3. TER3, TER4, SEX, HV_DOLLARS, LOW and GIFA did not contribute to any splits. Utilizing the confusion matrix for this model, as shown in Figure 4, the sensitivity, specificity, and overall accuracy were calculated to be 90.6%, 84.6%, and 87.6%, respectively.

Figure 4

JMP Output - Classification Decision Tree

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.5860	0.5260	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7416	0.6903	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2870	0.3285	∑ -Log(ρ[j])/n
RASE	0.2942	0.3137	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1748	0.1888	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1135	0.1244	∑ (ρ[j]≠ρMax)/n
N	3984	2018	n

Confusion Matrix

Training

Actual	Predicted Count	
donr	0	1
0	1740	249
1	203	1792

Validation

Actual	Predicted Count	
donr	0	1
0	862	157
1	94	905

Classification Model 2. Nominal Logistic Regression uses a function of the dependent variable, or logit function, to determine the probability of a binary classification outcome (Shmueli et. al, 2017). This method is popular, powerful, and inexpensive to run but may experience overfitting if the model is too complex. To run this initial model, we utilized all predictive variables except for KIDS, HV_DOLLARS and LOW. This model does require no multicollinearity which is why HV_DOLLARS and LOW were not used. Once the initial model was run, non-significant variables were removed from the effect summary. The final model excluded TER4, TER3, GIFA and SEX. The JMP fit details output is shown in Figure 4. For this model, the following measures of accuracy were calculated from the confusion matrix shown in Figure 5: sensitivity of 90.2%, specificity of 87.1%, and overall accuracy was 88.6%.

Figure 5

JMP Output - Nominal Logistic Regression

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.6551	0.6105	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7956	0.7613	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2391	0.2700	∑ -Log(ρ[j])/n
RASE	0.2729	0.2888	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1483	0.1597	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1034	0.1140	∑ (ρ[j]≠ρMax)/n
N	3984	2018	n

Lack Of Fit

Parameter Estimates

Confusion Matrix

Training

Actual	Predicted Count	
donr	0	1
0	1764	225
1	187	1808

Validation

Actual	Predicted Count	
donr	0	1
0	887	132
1	98	901

Classification Model 3. The Naïve Bayes classifier uses Bayes conditional probabilities to predict donors and non-donors based on the assumption that the predictors are independent of each other. The advantage of using Naïve Bayes is that it is a simple and computationally efficient model that can perform extremely well even if the underlying assumption of independent predictor variables is violated (Shmueli et. al, 2017). However, this model does require a substantial number of records to obtain adequate results (Shmueli et. al, 2017).

Based on the assumption of independence, INCAVG_DOLLARS and LOW were excluded, as well as KIDS and the log-transformed numeric variables in the model. Figure 6 details the output for the model after selecting and applying the appropriate response variable, predictor variables, and test/validation partition. Using the confusion matrix from the output information, sensitivity, specificity, and overall accuracy were calculated as 91.3%, 82.5% and 86.9%, respectively.

Figure 6

JMP Output - Naïve Bayes Classification Model

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.5157	0.5066	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6810	0.6727	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3357	0.3420	$\sum$ -Log(ρ[j])/n
RASE	0.3122	0.3184	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2226	0.2295	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.1260	0.1313	$\sum$ (ρ[j]≠ρMax)/n
N	3984	2018	n

donr

Confusion Matrix

Training

Actual donr	Predicted Count 0	Predicted Count 1
0	1653	336
1	166	1829

Validation

Actual donr	Predicted Count 0	Predicted Count 1
0	841	178
1	87	912

Classification Model 4. *k*-Nearest Neighbor (*k*-NN) can be used for both classification and regression predictive problems (Shmueli et. al, 2017). However, due to its simplicity and lack of parametric assumptions, we chose to utilize this method for classification (Shmueli et. al, 2017). To run the model the following predictor variables were used: TER1-4, OWND, KIDSB, INC, SEX, WLTH, HV_DOLLARS, INCMED_DOLLARS, LOW, NPRO, MDON, LAG, and GIFA. The following accuracy measures were calculated from the confusion matrix detailed in Figure 7: sensitivity 94.9%, specificity 68.9%, and overall accuracy 81.8%.

Figure 7

JMP Output - k-Nearest Neighbor

Training

K	Count	Misclassification Rate	Misclassifications	
1	3984	0.23042	918	
2	3984	0.23444	934	
3	3984	0.19603	781	
4	3984	0.19503	777	
5	3984	0.19152	763	
6	3984	0.19378	772	
7	3984	0.18725	746	
8	3984	0.19001	757	
9	3984	0.18700	745	*
10	3984	0.19177	764	

Validation

K	Count	Misclassification Rate	Misclassifications	
1	2018	0.22002	444	
2	2018	0.24182	488	
3	2018	0.19822	400	
4	2018	0.19673	397	
5	2018	0.19078	385	
6	2018	0.19029	384	
7	2018	0.18335	370	
8	2018	0.18186	367	*
9	2018	0.18335	370	
10	2018	0.18434	372	

Confusion Matrix for Best K=8

Training

Actual	Predicted Count	
donr	0	1
0	1356	633
1	124	1871

Validation

Actual	Predicted Count	
donr	0	1
0	702	317
1	50	949

Prediction Model 1. The first prediction model ran was a stepwise linear regression model. Stepwise regression models are beneficial in that they allow exploration of which predictor variables seem to provide a good fit utilizing different criteria in JMP's stepwise control panel. Through an iterative process, the model's prediction performance can be improved by reducing the variance by removing unnecessary terms and thus optimizing the model to predict DAMT (Shmueli et. al, 2017).

Based on the assumptions of linearity, the model excluded LOG[HV_DOLLARS], LOG[LOW] because they did not meet that assumption. We also excluded IDS, KIDS, DONR, and all the untransformed numeric variables as they were irrelevant. The assumption of linear residuals was assessed by viewing the Residual Normal Quantile Plot (Figure 8). It appears that the residuals are not exactly linear, however this is permissible as there is only slight deviation, which can be expected in some cases, and not a large deviation to heavily affect the model. To address multicollinearity, the parameter estimates platform on JMP was assessed and we removed any variables with a VIF higher than 10, however, there were no such cases (Figure 9). Lastly, utilizing a plot of standardized residuals versus predicted values, homoscedasticity was met, meaning that the residuals are equally distributed (Figure 10).

Utilizing JMP's automated variable selection, multiple iterations were performed to yield a reduced model with the highest R^2, lowest AICc/BICc, and lowest RASE, while maintaining parsimony. To fulfill the conditions above, we utilized a model with forward step and minimum AICc stopping criteria. We then reduced that model by removing any terms that were not significant, resulting in the final terms in Figure 11. From the summary of fit, as shown in Figure 12, the R^2 value is 0.591 and RASE (RMSE) is 1.2437.

Figure 8

JMP Output - Residual Normal Quantile plot for Stepwise Regression model

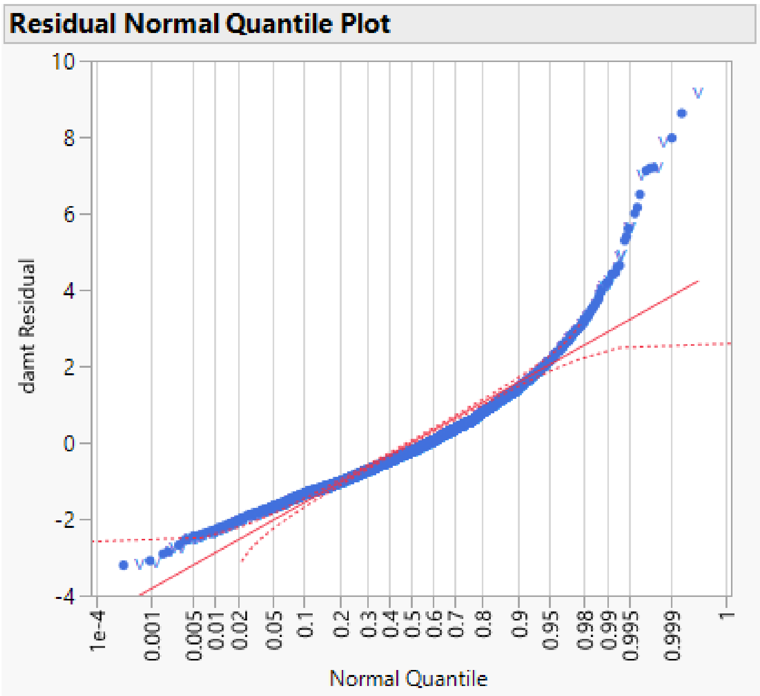


Figure 9

JMP Output - Parameter estimates of Stepwise Regression model

Parameter Estimates

Term	VIF
Intercept	.
ter3[0]	1.0515179
ter4[0]	1.0473761
ownd[0]	1.0380008
kidsb[1-0]	1.2034658
inc{1&2&3-4&5&6&7}	1.4640404
inc{1&2-3}	1.462736
inc{1-2}	1.7399491
inc{4-5&6&7}	1.1303886
sex[0]	1.0069098
wlth{0&1&2&3&4&5&6&7-8&9}	3.3731976
wlth{0&1&2&3-4&5&6&7}	3.5416457
wlth{0&1-2&3}	1.943857
wlth{2-3}	1.0154214
wlth{4&5-6&7}	1.1091515
Log[incmed_dollars]	1.0342762
Log[npro]	1.0258941
Log[gifa]	1.0065833

Figure 10

JMP Output - Standardized Residuals versus predicted values of DAMT for Stepwise Linear Regression model

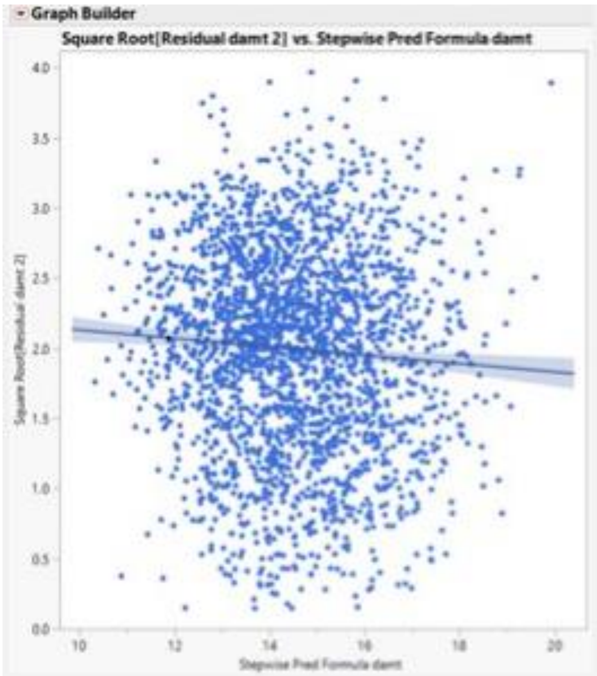


Figure 11

JMP Output - Effect Summary for Stepwise Linear Regression model.

Source	LogWorth		PValue
Log[gifa]	246.259		0.00000
ter4	80.212		0.00000
kidsb	58.396		0.00000
inc{1&2&3-4&5&6&7}	31.078		0.00000
ter3	25.526		0.00000
wlth{0&1&2&3-4&5&6&7}	19.686		0.00000
Log[npro]	9.471		0.00000
wlth{4&5-6&7}	9.317		0.00000
inc{4-5&6&7}	9.130		0.00000
Log[incmed_dollars]	7.510		0.00000
ownd	5.661		0.00000
wlth{0&1&2&3&4&5&6&7-8&9}	5.428		0.00000
inc{1&2-3}	4.950		0.00001
wlth{0&1-2&3}	2.769		0.00170
inc{1-2}	2.105		0.00785
sex	1.681		0.02085
wlth{2-3}	1.565		0.02725

Figure 12

JMP Output - Summary of Fit for the Stepwise Linear Regression model.

Summary of Fit	
RSquare	0.590911
RSquare Adj	0.587393
Root Mean Square Error	1.243667
Mean of Response	14.49825
Observations (or Sum Wgts)	1995

Prediction Model 2. The second prediction model explored was a regression tree. This model was viewed as a good type for our analysis, as it works best with large data sets and requires no statistical model assumptions. Regression trees "branch" the data into optimal subsets to maximize the differences between grouping means, thus creating optimally unique grouping.

To create our regression tree model, we set our Y-variable as DAMT, and included TER1, TER2, TER3, TER4, OWND, KIDS, INC, SEX, WLTH, HV_DOLLARS, INCMED_DOLLARS, LOW, NPRO, MDON, LAG, and GIFA as our predictor variables. The validation set was also utilized to evaluate the strength of the model. The resulting model had 32 splits, RASE of 1.5021, and an R^2 value of 0.476 indicating that approximately half of the DAMT variance is explained by the model (Figure 13). Ultimately, only ten of the original sixteen variables contributed to the prediction enough to be included in the final model, with GIFA accounting for the largest portion of the model (Figure 14).

Figure 13

JPM Output - Regression Tree Model

	RSquare	RASE	N	Number of Splits	AICc
Training	0.547	1.3031266	1995	32	6787.2
Validation	0.476	1.5021157	999		

Figure 14

JPM Column Contributions Output for Regression Tree Model

Column Contributions

Term	Number of Splits	SS	Portion
gifa	8	2437.30181	0.5964
ter4	4	740.145624	0.1811
kids	5	413.32744	0.1011
ter3	3	189.105582	0.0463
inc	5	156.385436	0.0383
wlth	3	80.0166761	0.0196
npro	1	25.9006092	0.0063
hv_dollars	1	15.3846154	0.0038
ter2	1	14.7740453	0.0036
ownd	1	14.6147533	0.0036
ter1	0	0	0.0000
sex	0	0	0.0000
incmed_dollars	0	0	0.0000
low	0	0	0.0000
mdon	0	0	0.0000
lag	0	0	0.0000

Prediction Model 3. Whereas the previous model is based on the results of a single tree, the next prediction model – the bootstrap forest regression tree – is based on the results of multiple trees to improve model performance. As in our initial regression tree, the bootstrap forest regression tree requires none of the statistical model assumptions, therefore we used the same input variables described above. The goal of the bootstrap forest regression tree is to increase the R^2 value and decrease the RASE in comparison to the single regression tree model, indicating improved prediction ability.

The model was based on a 100-tree forest, with four terms sampled per split at a sample rate of one. Random seed settings were left at zero and early stopping was allowed. The resulting model had a RASE of 1.4265 (single regression tree model RASE=1.5021) and an R^2 of 0.528 (single regression tree model R^2=0.476). These numbers indicate a slight improvement in predictive ability using the bootstrap forest regression tree model (Figure 15). All sixteen of the variables were left in the model, with GIFA still contributing the largest portion to the model prediction (Figure 16).

Figure 15

JPM Output – Bootstrap Forest Regression Tree Statistics

Overall Statistics

Individual Trees	RASE
In Bag	0.974120
Out of Bag	1.676545

	RSquare	RASE	N
Training	0.683	1.0894956	1995
Validation	0.528	1.4264502	999

Figure 16

JPM Output – Column Contributions for Bootstrap Forest Regression Tree

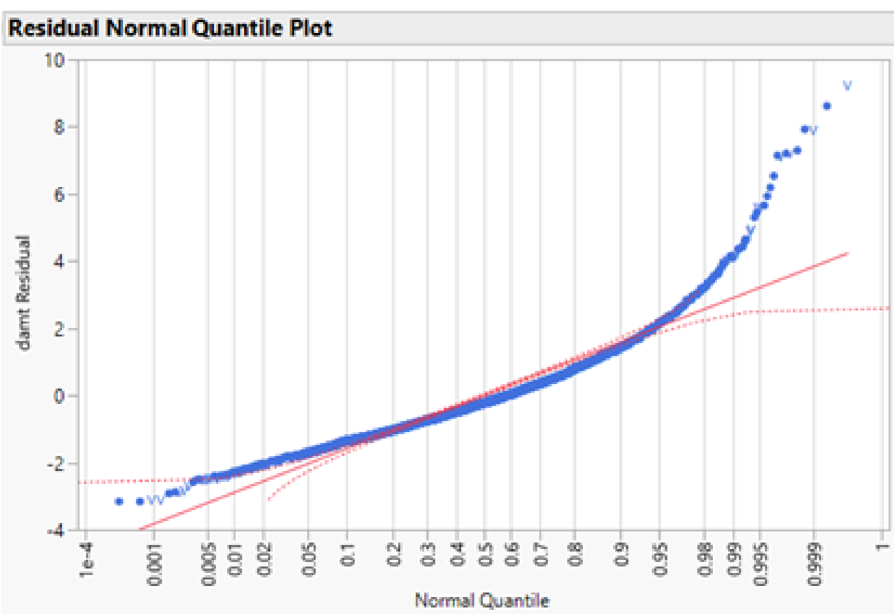
Column Contributions

Term	Number of Splits	SS	Portion
gifa	2313	1204.25439	0.4275
ter4	279	335.049426	0.1189
kids	1019	235.143209	0.0835
inc	929	126.171894	0.0448
wlth	1191	116.429723	0.0413
npro	1433	106.848895	0.0379
ter2	504	106.659514	0.0379
low	1437	102.221091	0.0363
incmed_dollars	1413	98.7817299	0.0351
hv_dollars	1404	98.682827	0.0350
ter3	226	77.3018544	0.0274
mdon	1261	75.9364973	0.0270
lag	1142	58.6949027	0.0208
ter1	433	43.8285394	0.0156
sex	823	23.5587824	0.0084
ownd	71	7.7083114	0.0027

Prediction Model 4. Standard Least Squared model is a Linear Regression used in the fourth prediction model, and we attempted to model the relationship between several explanatory variables (KIDSB, INC, OWND, WLTH, LOG[INCMED_DOLLARS], LOG[NPRO], TER3, LOG[GIFA], TER4, and SEX, and our response variable, DAMT. Once the model was run, we examined a graph Residuals by Predicted DAMT at DONR = 1, which shown no distinct pattern with seemingly constant variance, indicating a linear relationship and meeting the assumption of linearity required (Shmueli et. al, 2017). A visual inspection at a graph built with Residuals from the model against the predicted values for DAMT, affirms that the condition of homoscedasticity is met, as no distinguishable funnel shape is recognized, meaning that variation is considered constant around the regression line (Shmueli et. al, 2017). By transforming predictors in this model, the Residual Normal Quantile Plot shows a close enough normal behavior of the variables, and even though the residuals show some deviation, we will consider it as normal since there is no drastic deviations (Figure 17).

Figure 17

JMP Output – Residual Normal Quantile Plot



This model had a RMSE (RASE) value of 1.2434, R^2 value of 0.592, placing it among the good candidates to explain over half of the DAMT variance, also it possessed low AIC of 6557.4 and BIC of 6696.7, which are competitive to measure the goodness of fit (Shmueli et. al, 2017) among the other models considered (see Figure 18).

Figure 18

JMP Output – Summary of Fit

Summary of Fit

RSquare	0.592304
RSquare Adj	0.587547
Root Mean Square Error	1.243434
Mean of Response	14.49825
Observations (or Sum Wgts)	1995

AICc	BIC
6557.409	6696.709

To determine the best classification model, we evaluated the models based on the resulting return on interest (ROI) rate (Net Funds Gained / Capital Investment × 100) and measures of accuracy. Cost and benefits matrices, as shown in Figure 19, were created using the confusion matrixes of our classification models. According to our figures, the Nominal Logistic Regression model would be the most suitable classifier as it results in the highest rate of ROI and overall accuracy as well with a value of 88.6%. While the sensitivity rate is lowest, suggesting a higher number of true donors classified as non-donors, the specificity rate is highest which means the least number of non-donors were wrongly classified as donors. These rates are important to consider when the goal is to maximize profits and minimize costs. We want to minimize false positives as this would lead to less profit per mailing. For example, k-NN appears to have the highest net profit, however, this model misclassified more non-donors as donors, which increases mailing costs and lowers the specificity (Figure 19). This in turn decreases the rate of ROI as more capital needs to be invested. Therefore, the Nominal Logistic Regression model is the best classification model as it is the most accurate and results in the highest rate of ROI (532%).

Figure 19

Cost and benefits matrices for classification models

Confusion Matrices

Decision Tree (Validation)		
Actual	Predicted	
donr	0	1
0	862	157
1	94	905

Nominal Logistic Regression (Validation)		
Actual	Predicted	
donr	0	1
0	887	132
1	98	901

Naïve Bayes (Validation)		
Actual	Predicted	
donr	0	1
0	841	178
1	87	912

k-NN (Validation)		
Actual	Predicted	
donr	0	1
0	702	317
1	50	949

Accuracy Measures

Decision Tree Validation	
Sensitivity	0.905906
Specificity	0.845927
Misclassification rate	12.44%
Overall accuracy	87.56%

Nominal Logistic Regression	
Sensitivity	0.901902
Specificity	0.870461
Misclassification rate	11.40%
Overall accuracy	**88.60%**

Naïve Bayes	
Sensitivity	0.912913
Specificity	0.825319
Misclassification rate	13.13%
Overall accuracy	86.87%

k-NN	
Sensitivity	0.94995
Specificity	0.688911
Misclassification rate	18.19%
Overall accuracy	81.81%

Profit Matrices

Decision Tree		
Profit	Predicted	
donr	0	1
0	0	0
1	0	$13,122.50
	Net:	**$10,998.50**

Nominal Logistic Regression		
Profit	Predicted	
donr	0	1
0	0	0
1	0	$13,064.50
	Net:	**$10,998.50**

Naïve Bayes		
Profit	Predicted	
donr	0	1
0	0	0
1	0	$13,224.00
	Net:	**$11,044.00**

k-NN		
Profit	Predicted	
donr	0	1
0	0	0
1	0	$13,760.50
	Net:	**$11,228.50**

Cost Matrices

Decision Tree		
Cost	Predicted	
donr	0	1
0	$0.00	-$314.00
1	-$1,363.00	-$1,810.00
ROI Rate		518%
Profit/mailing		$10.36

Nominal Logistic Regression		
Cost	Predicted	
donr	0	1
0	$0.00	-$264.00
1	-$1,421.00	-$1,802.00
ROI Rate		532%
Profit/mailing		$10.65

Naïve Bayes		
Cost	Predicted	
donr	0	1
0	$0.00	-$356.00
1	-$1,261.50	-$1,824.00
ROI Rate		507%
Profit/mailing		$10.13

k-NN		
Cost	Predicted	
donr	0	1
0	$0.00	-$634.00
1	-$725.00	-$1,898.00
ROI Rate		443%
Profit/mailing		$8.87

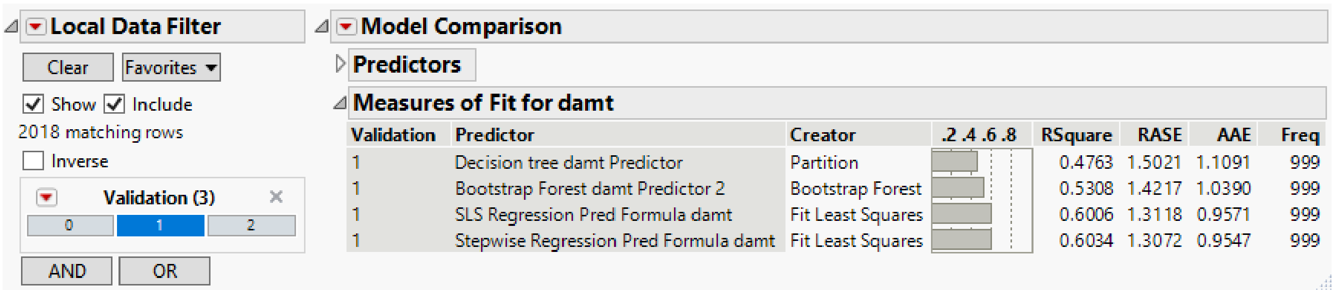
Key:

- Donors
- Misclassified Non-Donors
- Gross Profits

To determine the best prediction model, we evaluated the models based on the RASE (RMSE) criteria as this measures the accuracy of the prediction models (Shmueli et. al, 2017). The lower the RASE, the more accurate the predictive capabilities of that model are. Therefore, by looking at the model comparison output in Figure 20 for the validation data, the model with the lowest RASE is the Stepwise Linear Regression model.

Figure 20

JMP Output - Model Comparison of all predictive models for donors only



Conclusion

Evaluation

To recapture the intent of this study, we are utilizing a classification model and a prediction model to 1) improve cost-effectiveness of direct marketing campaigns by accurately classifying which donors are more likely to respond to the mailing campaign with a new donation, 2) predict the donation amounts of said donors, and 3) maximize return on investment. By leveraging the resulting predictions from the two models, net proceeds and ultimately, ROI can be maximized.

As with any model, there are limitations and possible sources of bias of the results generated. For both of our models, the conditions that were required to be met before running the models posed as a part of those limitations and bias. For our linear prediction model, assumptions of linearity, little/no multicollinearity, independent errors, equal variance (homoscedasticity), and normally distributed errors needed to be met (Field, 2018). For our logistic classification model, the assumption of linearity between the predictors and the logit of the response variable, independence of residuals, and no multicollinearity needs to be met (Shmueli et. al, 2017). Depending on the nature of the data and the business objectives, the approach on how to address these assumptions, if violated, can vary. However, for our study, if an assumption was not met then variable transformation and/or removal of variables was done.

To improve cost-effectiveness and obtain maximum net proceeds and ROI, we first want to accurately target responders in our sample who are most likely to donate. Secondly, we need to accurately predict the expected donation amounts from our likely donors. Therefore, we recommend applying the Nominal Logistic Regression classification model as it accurately classifies donors at a rate of 90.19% and non-donors at a rate of 87.05% and has the highest overall accuracy of 88.60%. These rates are important to consider as they affect net profits and capital investment and ultimately ROI. We also recommend applying the Stepwise Regression Model as it has the highest predictive ability of expected gift amounts from likely donors (Figure 20).

By combining the test dataset predictions from the logistic and stepwise regression models, the organization’s capital investment would be approximately $832.00, and the estimated net proceeds would be $5096.92 after receiving an estimated average of $14.25 per donor, as shown in Table 5. While this average is lower than $14.50, as stated in the beginning, the lift indicates that the nominal logistic regression model correctly identifies approximately two times the number of donors than random sampling when looking at the top 10% (portion 0.1) of a new sample of the same size (Figure 21). This means that a direct marketing campaign utilizing these models would reach those who are likely to donate, thus maximizing net proceeds by keeping costs low. To conclude, maximizing ROI can be achieved by utilizing our models as they specifically target classified donors at twice the original rate, minimize costs, and have the highest predictive performance.

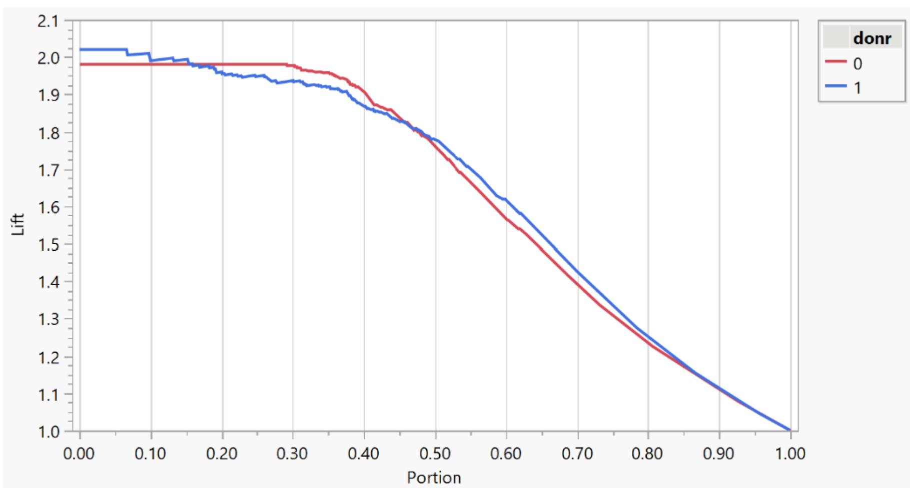
Table 5

Test dataset predictions from Stepwise Regression and Nominal Logistic Regression models

Logistical Regression * Stepwise Regression		
Donor QTY per model	N	416
Donation total estimated	$	5927.92
Inv 2 USD / mailed donor	$	832
Estimated Average Donation	$	14.25
Net profit / model	$	5095.92

Figure 21

JMP Output - Lift curve for Nominal Logistic Regression model



References

Field, Andy. (2018). *Discovering Statistics Using IBM SPSS Statistics*. (5th Edition). SAGE Publications.

Shmueli, G., Bruce, P. & Stephens, M. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro*. John Wiley & Sons, Inc.

What is a "Nonprofit"?. (2020). National Council of Nonprofits. https://www.councilofnonprofits.org/what-is-a-nonprofit

Appendix

Figure A1

Bar chart of observations by home ownership status (stacked by response type)

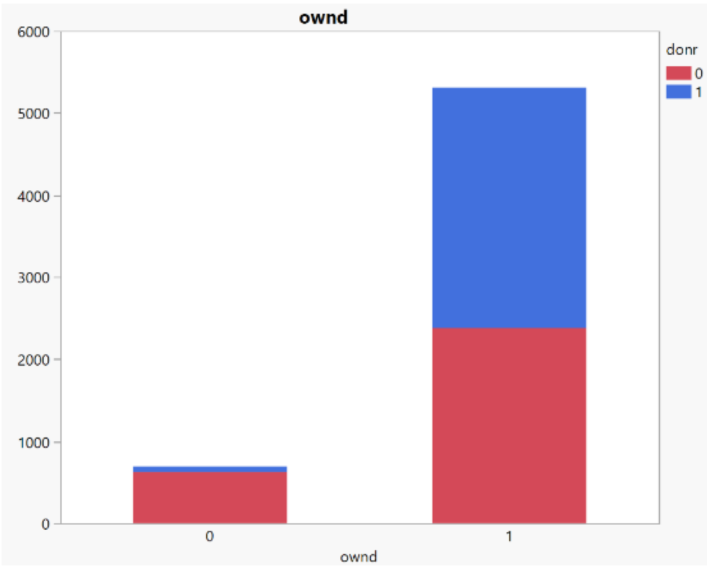


Figure A2

Box chart for donation amount by home ownership status (stacked by response type)

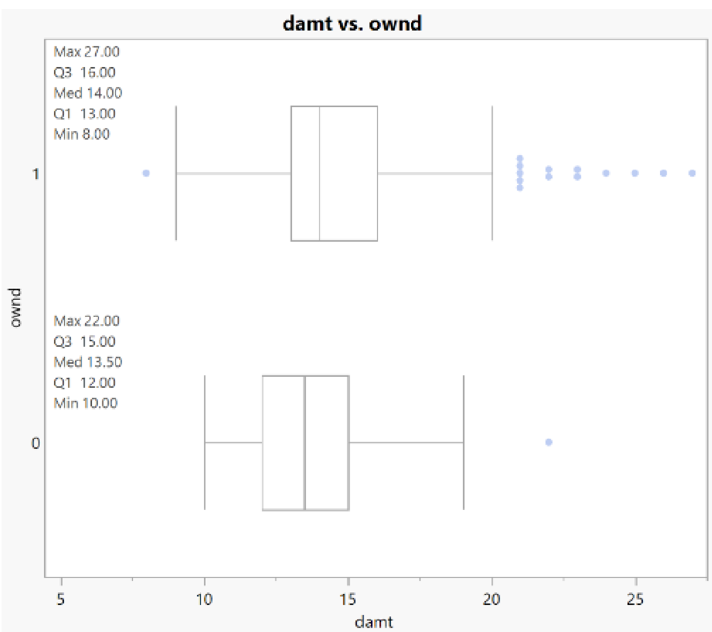


Figure A3

Bar chart of observations by number of children (stacked by response type)

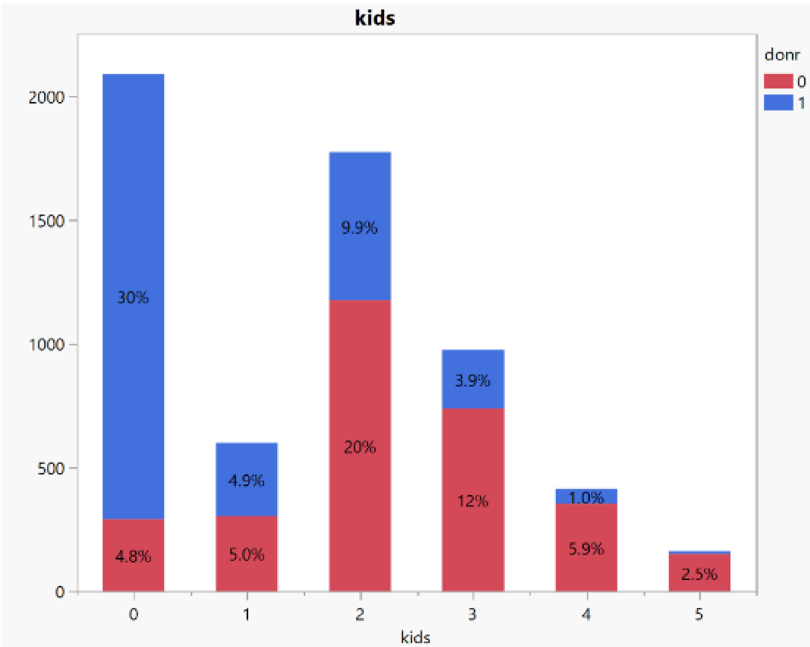


Figure A4

Box chart for donation amount by number of children

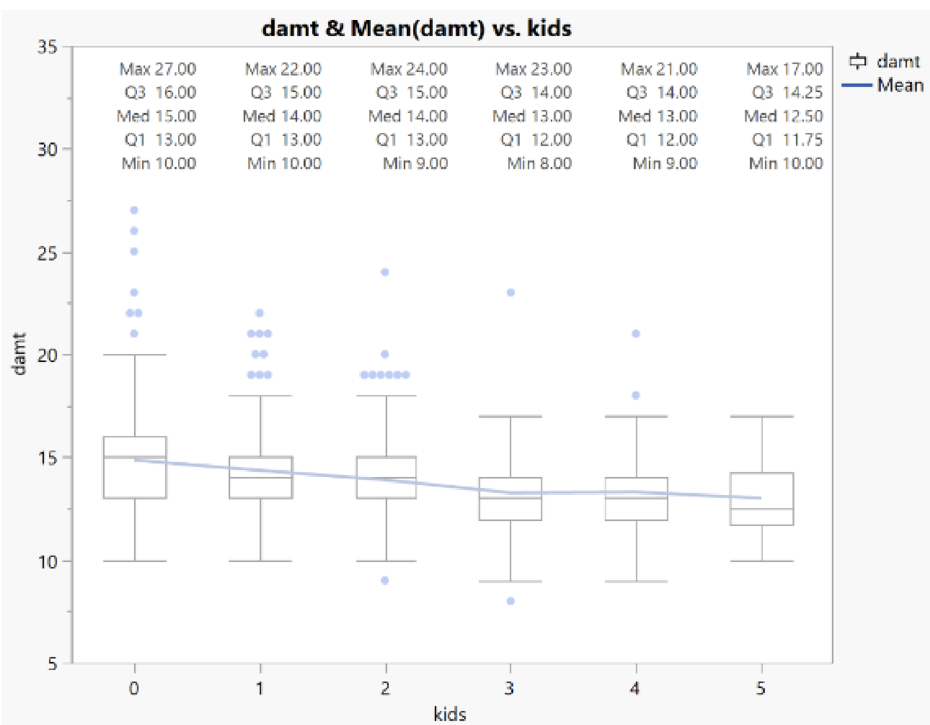


Figure A5

Bar chart of observations by household income level (stacked by response type)

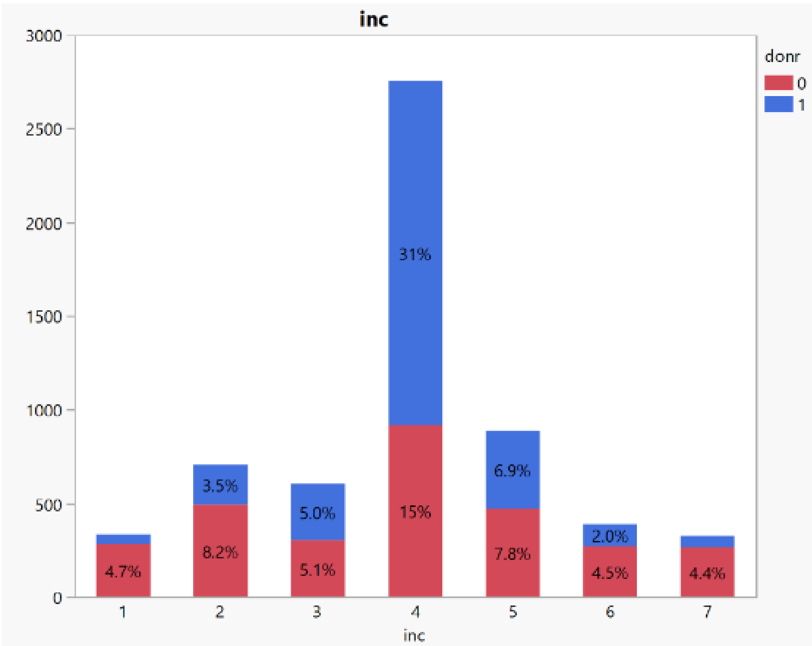


Figure A6

Box chart for donation amount by household income level

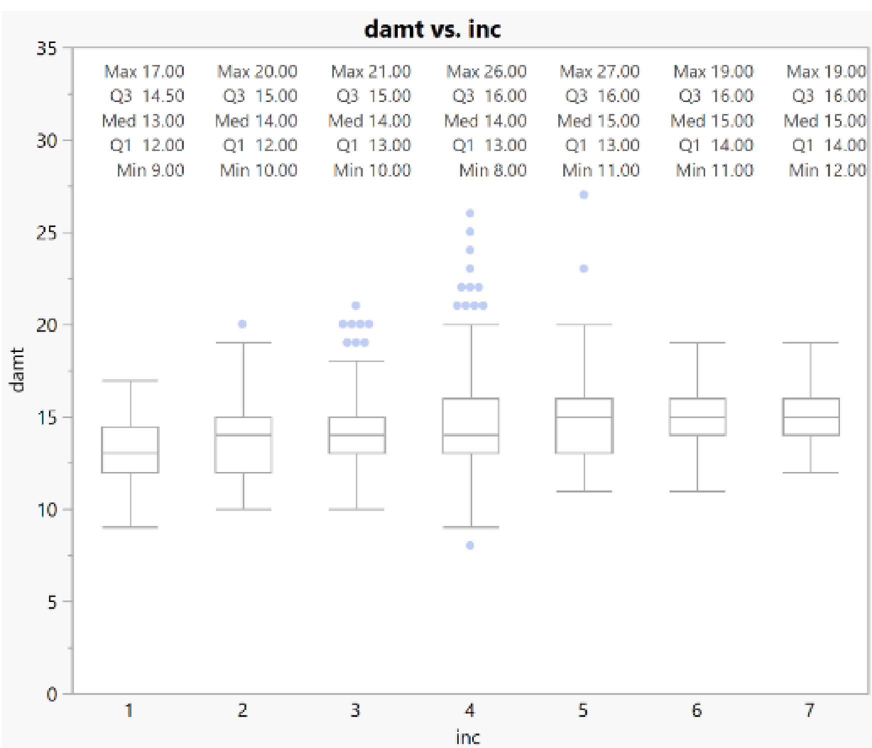


Figure A7

Bar chart of observations by gender (stacked by response type)

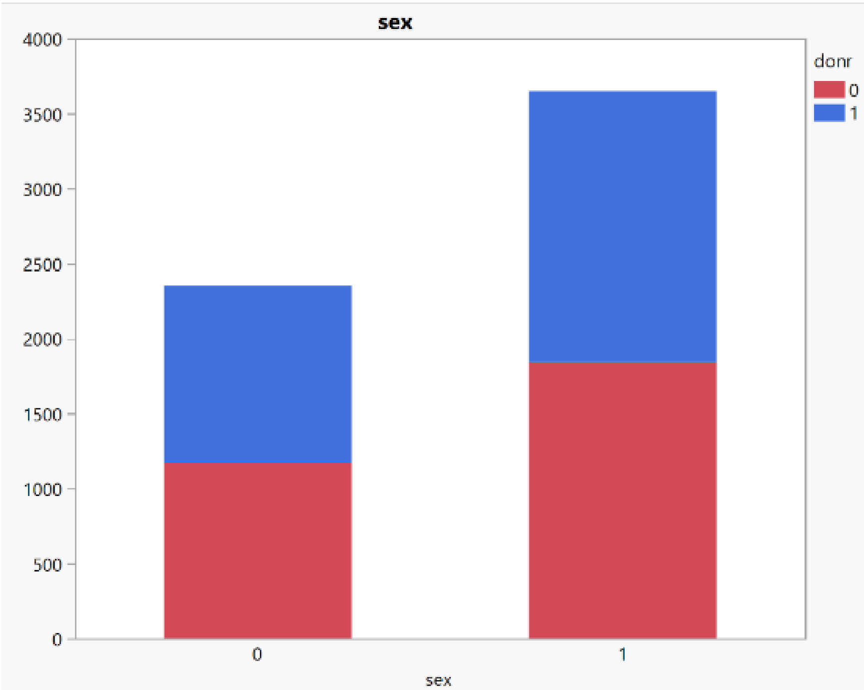


Figure A8

Box chart for donation amount by gender

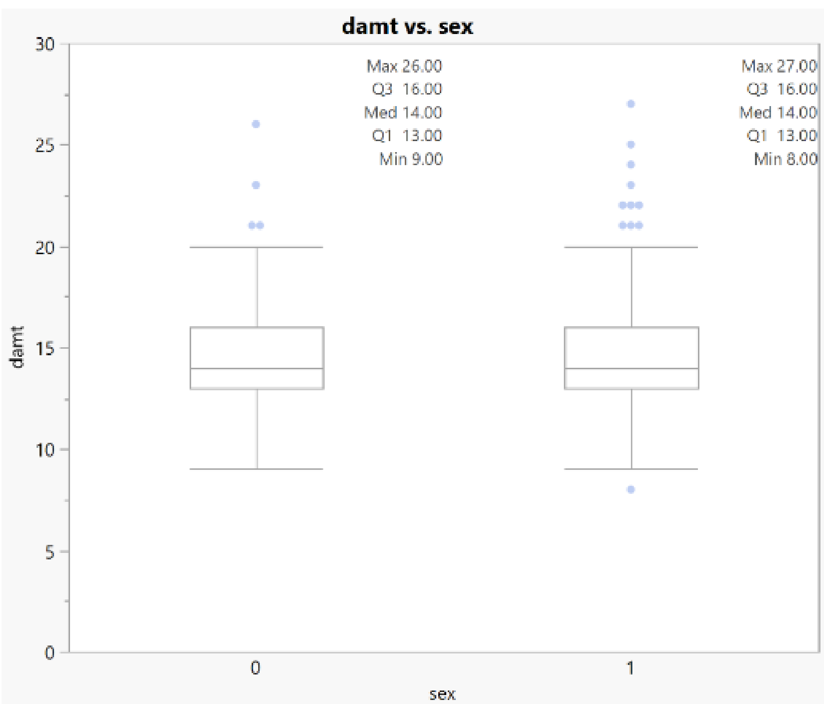


Figure A9

Histogram of observations by wealth rating (stacked by response type)

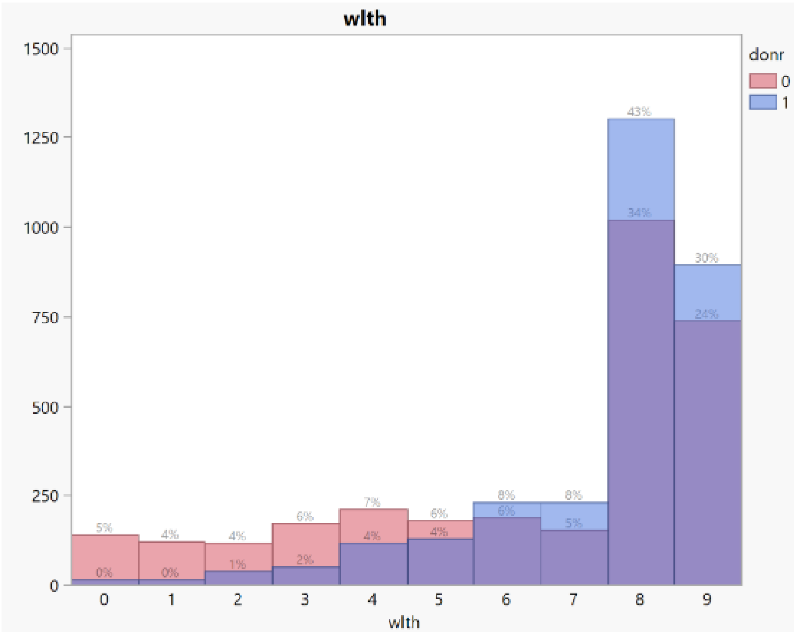


Figure A10

Box chart for donation amount by wealth rating

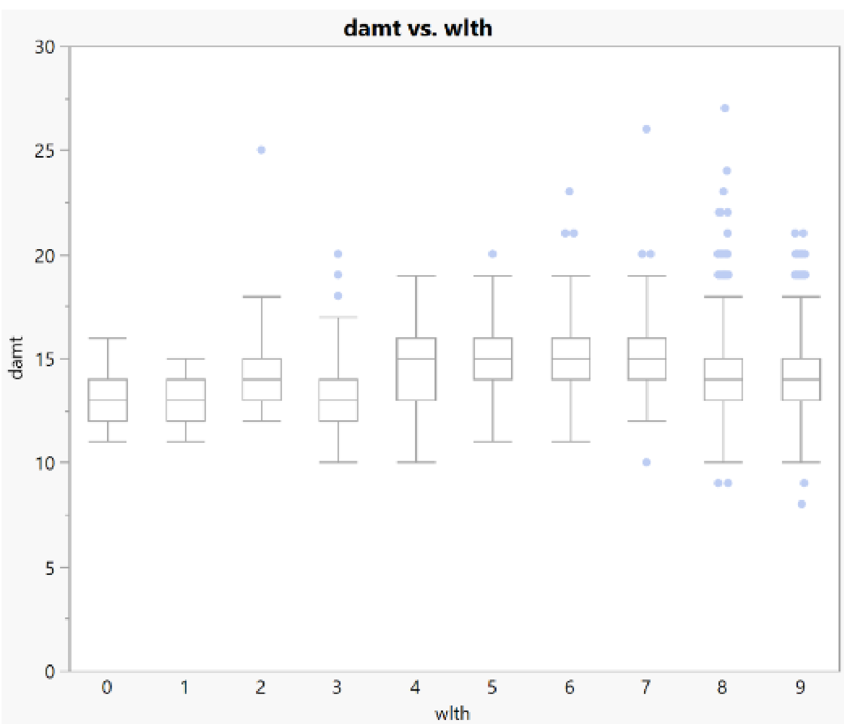


Figure A11

Histogram of observations by neighborhood home value (stacked by response type)

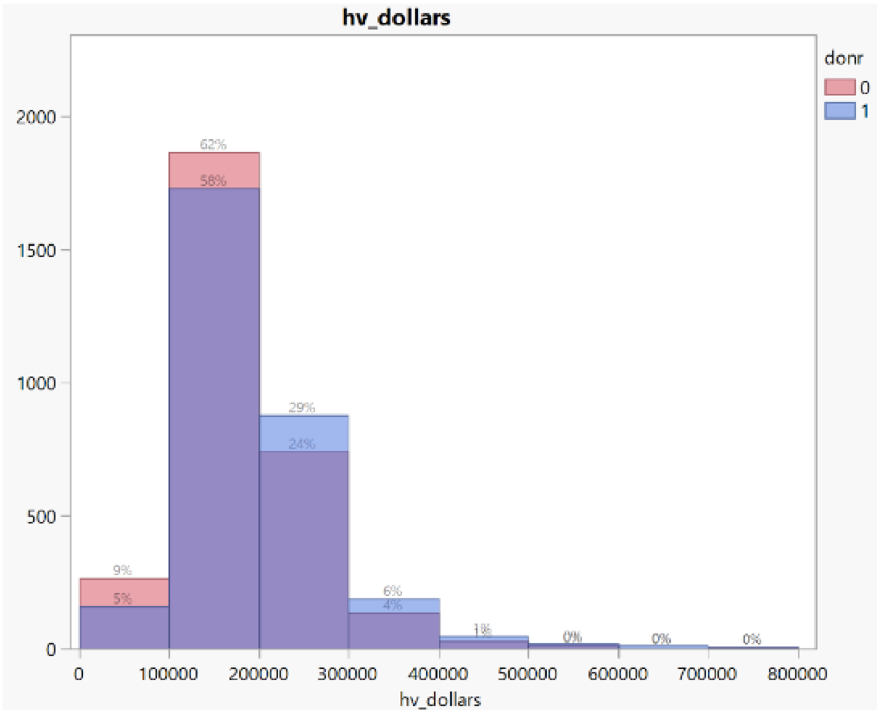


Figure A12

Bivariate fit of donation amount by neighborhood home value

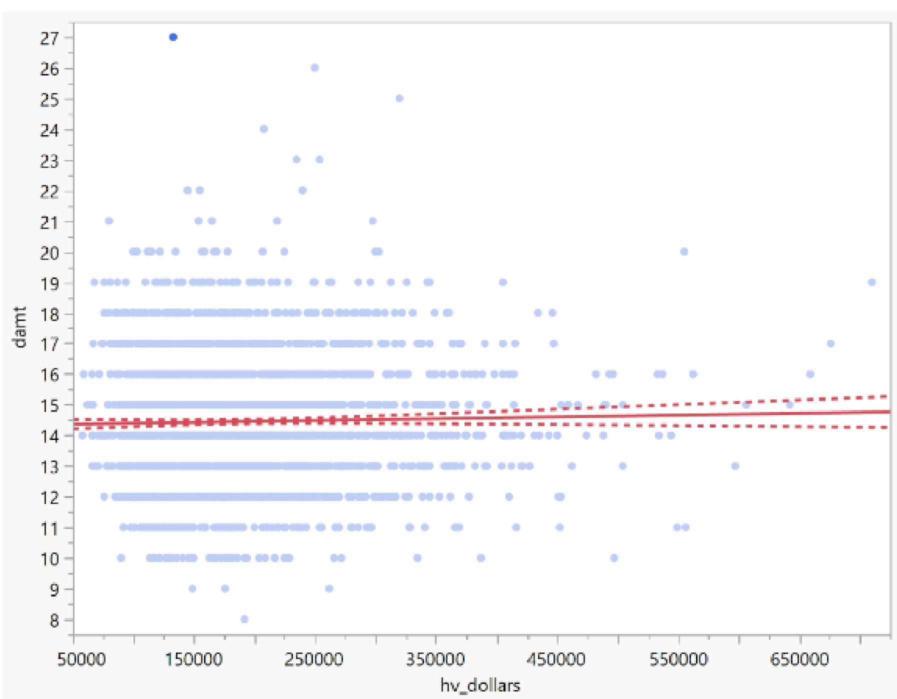


Figure A13

Histogram of observations by median neighborhood income (stacked by response type)

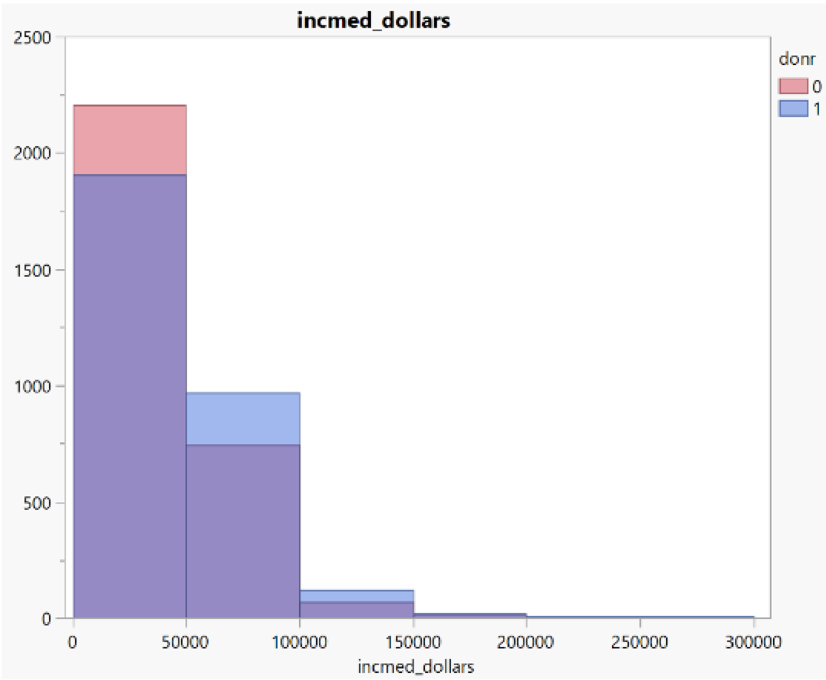


Figure A14

Bivariate fit of donation amount by median neighborhood income

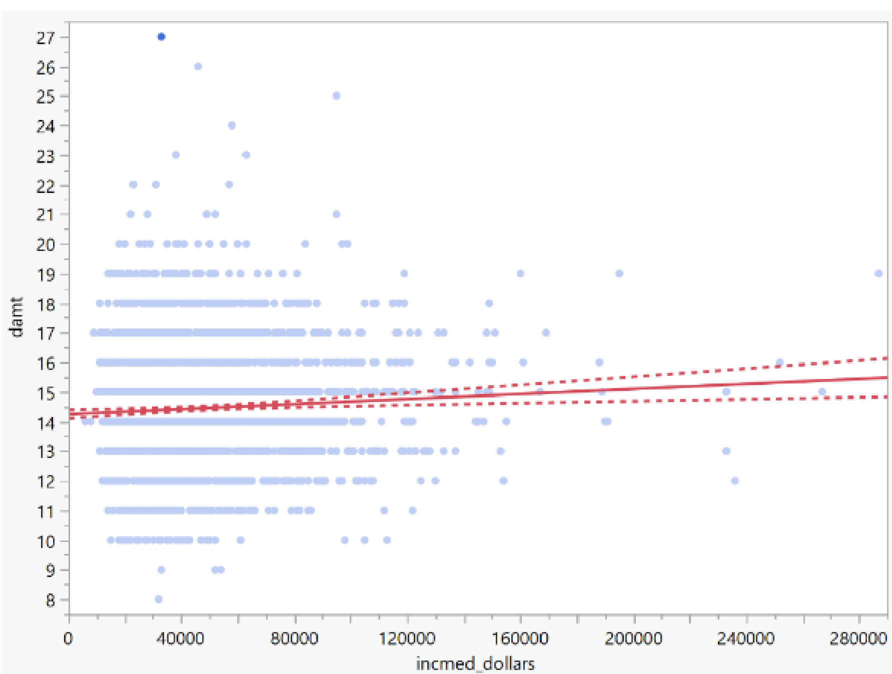


Figure A15

Histogram of observations by neighborhood percentage low income (stacked by response type)

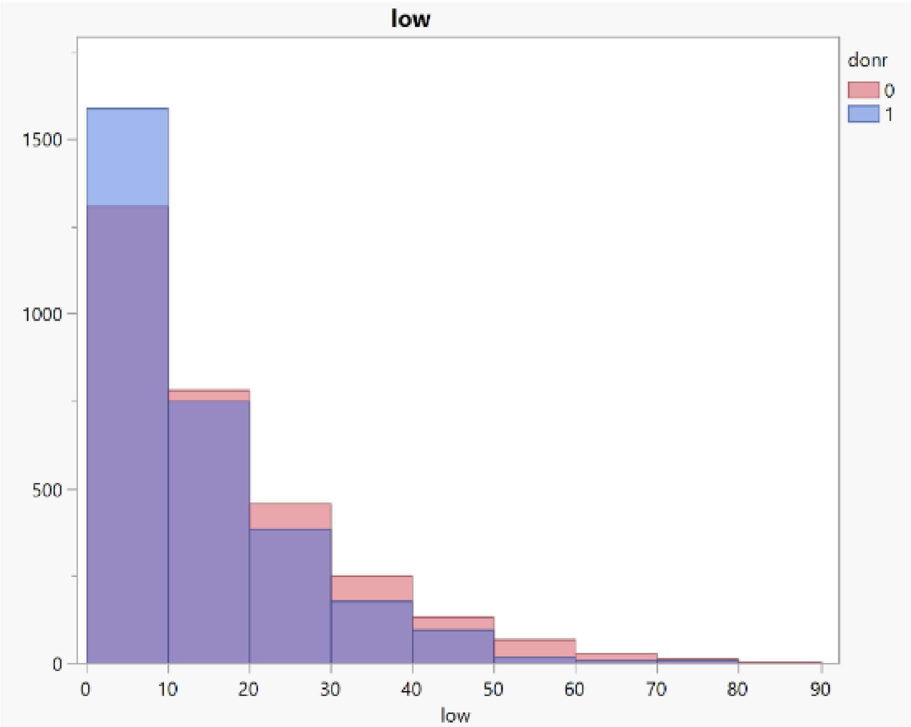


Figure A16

Bivariate fit of donation amount by neighborhood percentage low income

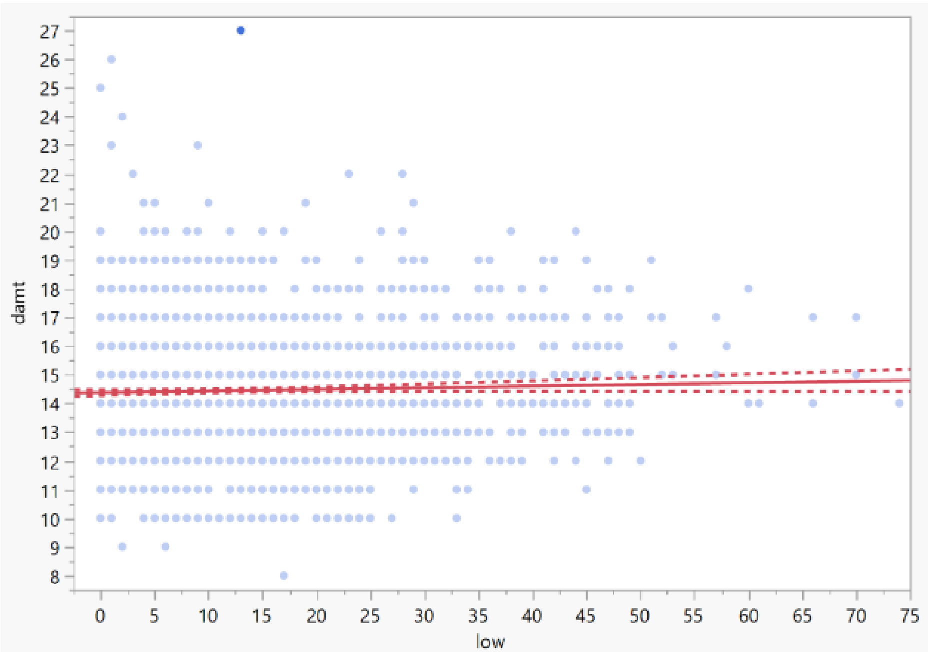


Figure A17

Histogram of observations by total promotions received (stacked by response type)

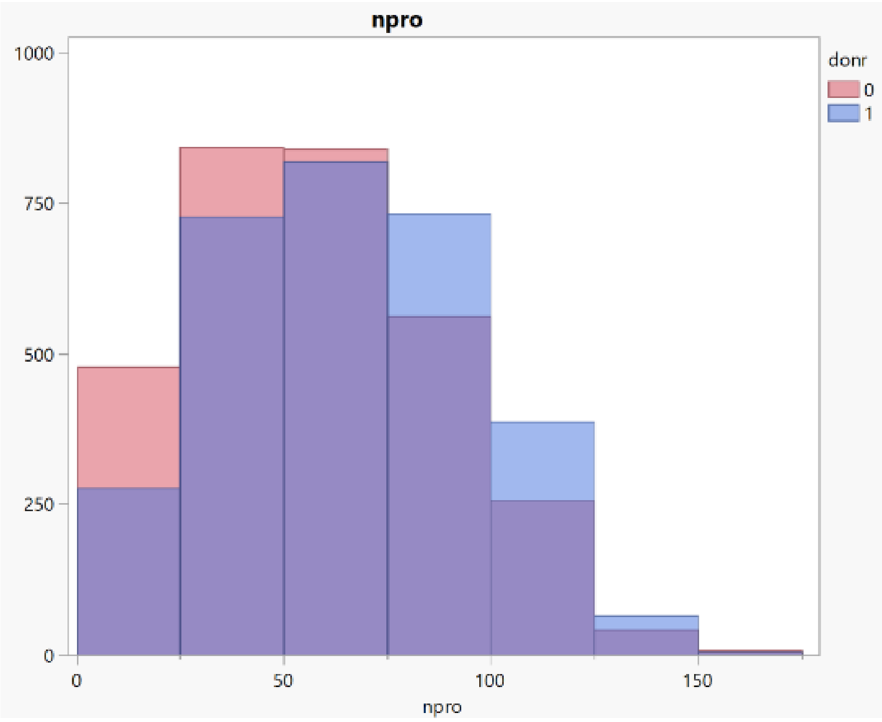


Figure A18

Bivariate fit of donation amount by total promotions received

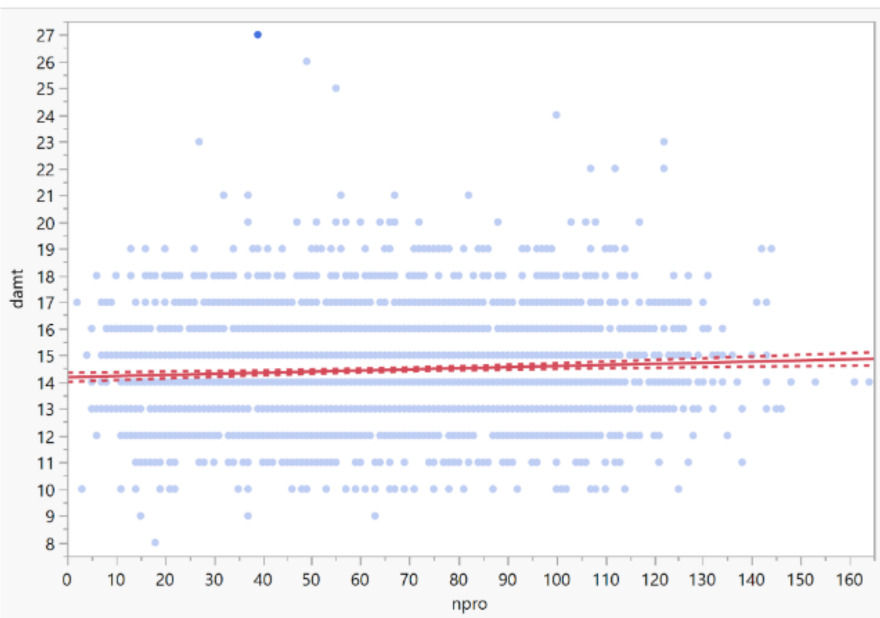


Figure A19

Histogram of observations by months since last donation (stacked by response type)

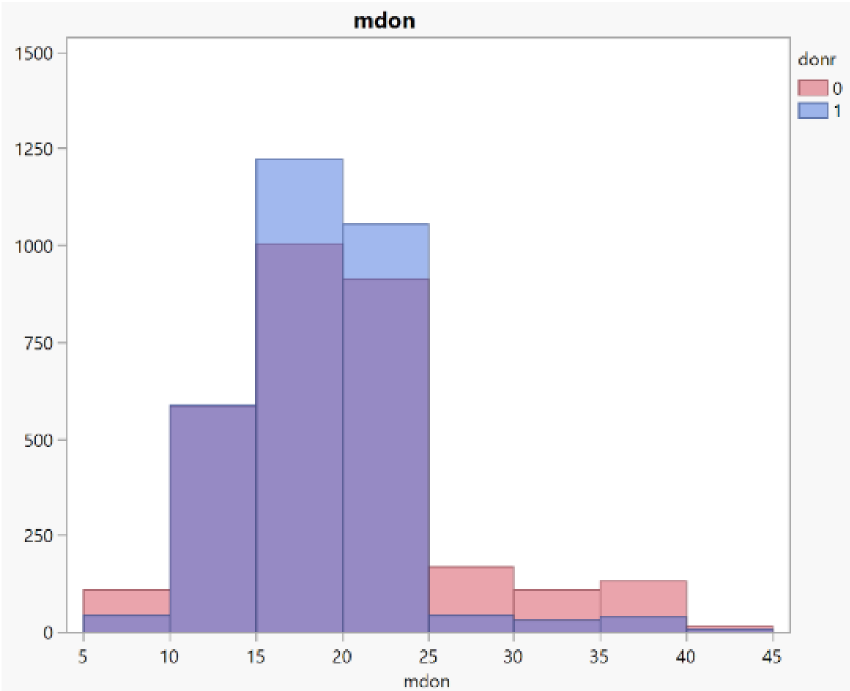


Figure A20

Bivariate fit of donation amount by months since last donation

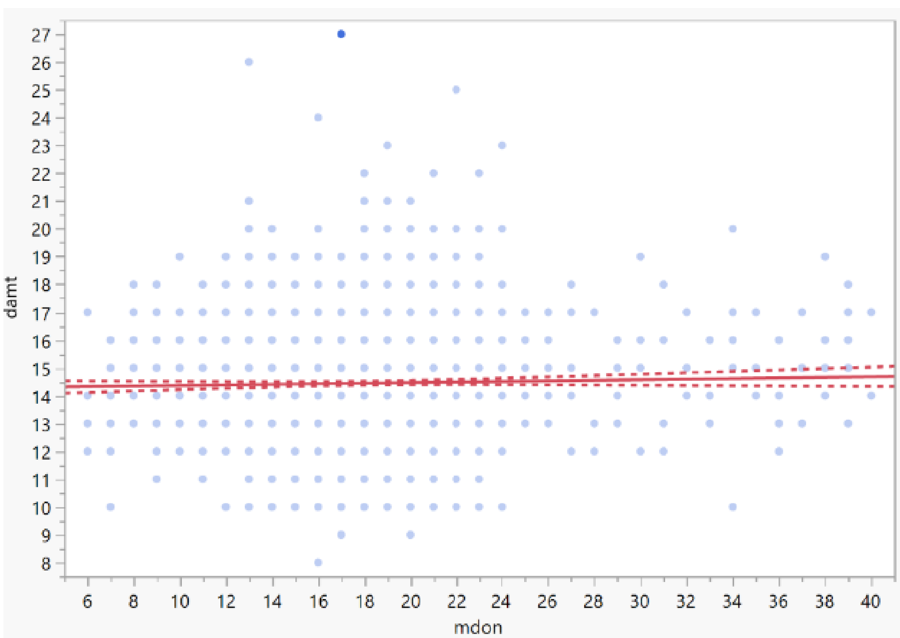


Figure A21

Histogram of observations by months between first and second donation (stacked by response type)

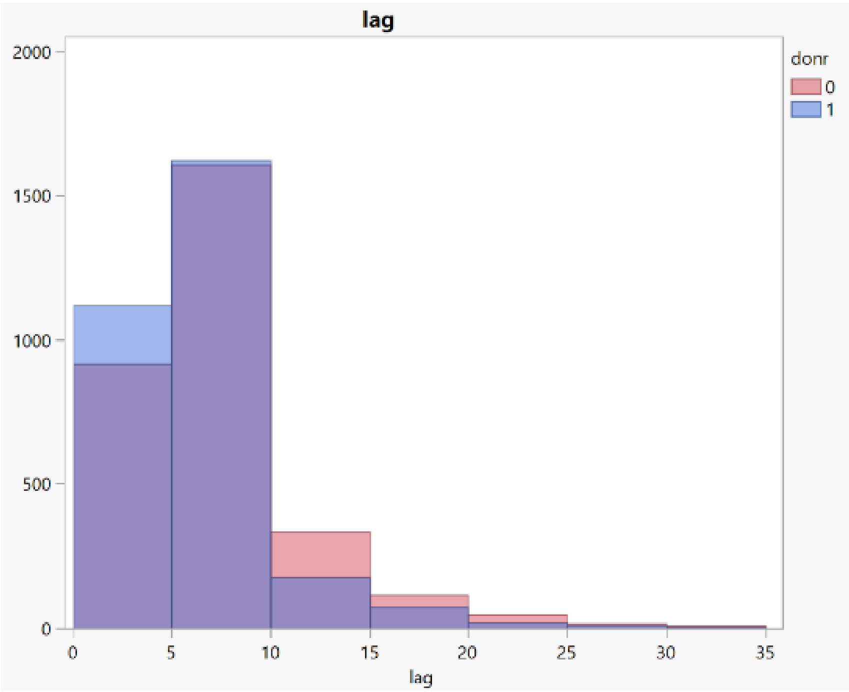


Figure A22

Bivariate fit of donation amount by months between first and second donation

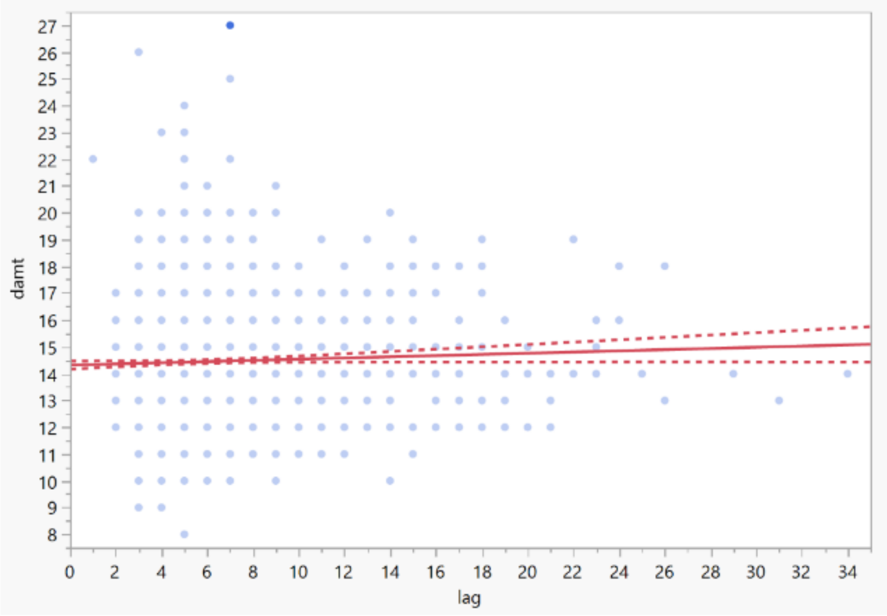


Figure A23

Histogram of observations by average value of gifts received (stacked by response type)

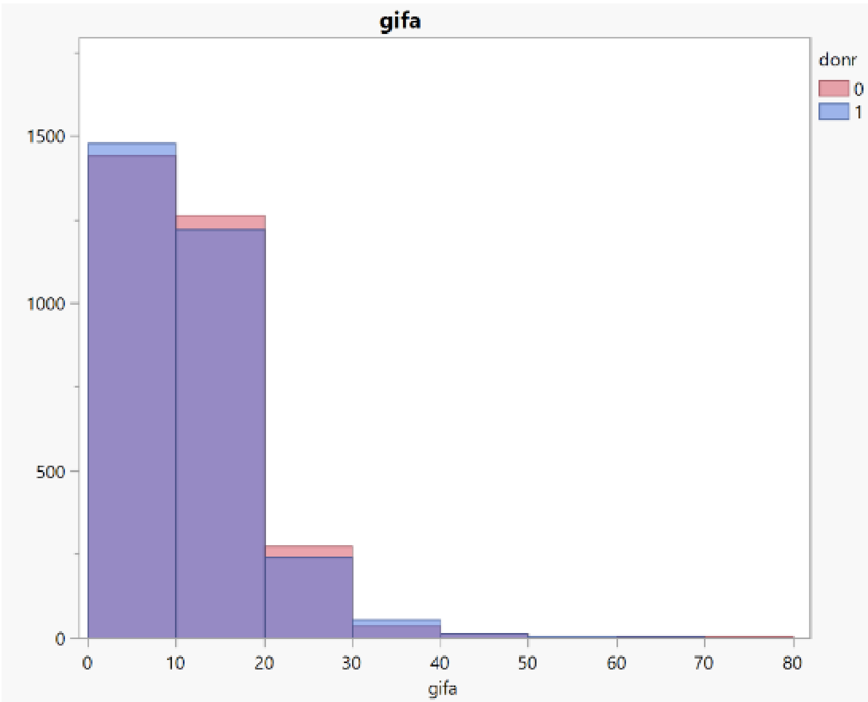


Figure A24

Bivariate fit of donation amount by average value of gifts received

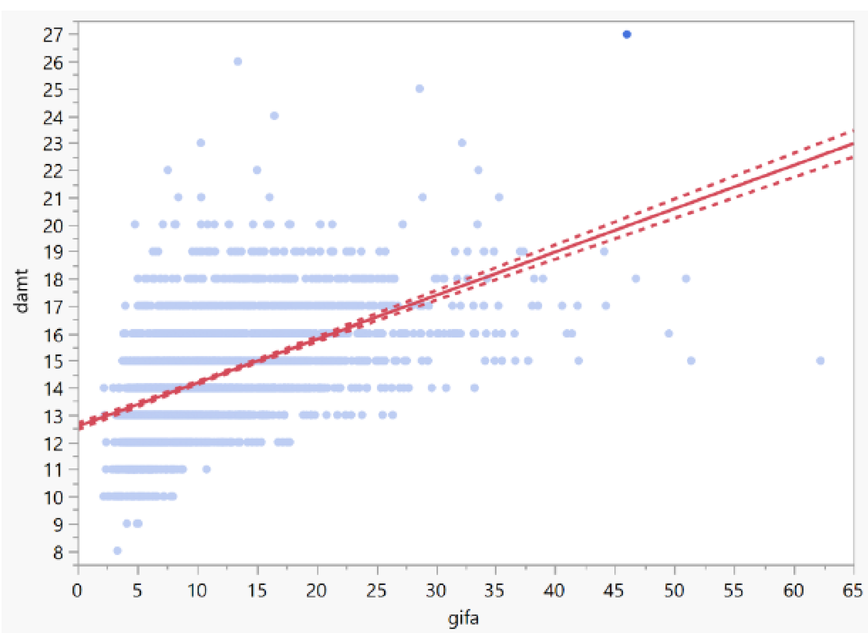


Figure A25

Pie chart of observations by region

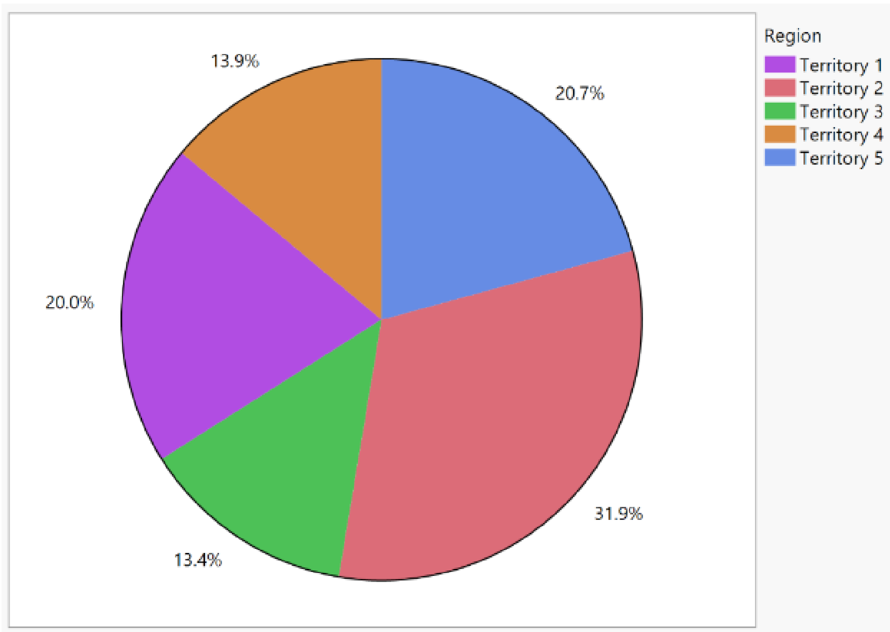


Figure A26

Response rate by region (5 – 1 from left to right)